Humans, machines, and language: A deep alignment in underlying computational styles?

Bingjiang Lyu*, Lorraine K. Tyler*, Yuxing Fang, William D. Marslen-Wilson

Centre for Speech, Language and the Brain, Department of Psychology,

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

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

Abstract

The emergence of AI systems that emulate the remarkable human capacity for language has raised fundamental questions about complex cognition in humans and machines. This lively debate has largely taken place, however, in the absence of specific empirical evidence about how the internal operations of artificial neural networks (ANNs) relate to processes in the human brain as listeners speak and understand language. To directly evaluate these parallels, we extracted multi-level measures of word-by-word sentence interpretation from ANNs, and used Representational Similarity Analysis (RSA) to test these against the representational geometries of real-time brain activity for the same sentences heard by human listeners. These uniquely spatiotemporally specific comparisons reveal deep commonalities in the use of multi-dimensional probabilistic constraints to drive incremental interpretation processes in both humans and machines. But at the same time they demonstrate profound differences in the underlying functional architectures that implement this shared algorithmic alignment.

(147 words)

Keywords: computational cognitive neuroscience, artificial neural networks, incremental language comprehension, natural language processing, Representational Similarity Analysis, EEG/MEG

Introduction

Modern artificial neural networks (ANNs) have made great strides in achieving human-like levels of performance in tasks such as visual object recognition (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015) and speech recognition (Hinton et al., 2012) as well as in more complex domains such as natural language processing (NLP) (Brown et al., 2020; Devlin et al., 2018) that reflect unique and foundational human cognitive capacities. But despite these growing parallels between the capacities of humans and ANNs in higher order task domains, it remains unclear what is the specific relationship between the computational principles and representations implemented within the very different architectures of brains and ANNs (Lillicrap et al., 2020). Without a real understanding of this relationship, any attempt to evaluate the significance of these developments will not be empirically well-founded. This critical question also speaks to whether and how we can use ANNs to better understand and even reverse-engineer the neurocomputational mechanisms underpinning human cognition and intelligence (Cichy and Kaiser, 2019; Devereux et al., 2018; Kriegeskorte, 2015; Richards et al., 2019; Saxe et al., 2021; Yamins and DiCarlo, 2016; Yang and Wang, 2020).

Recent studies have suggested potential commonalities between brains and ANNs in a wide array of cognitive functions (Caucheteux and King, 2020; Goldstein et al., 2021; Guclu and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2021; Schrimpf et al., 2020; Sheahan et al., 2021), but largely do so by probing brain activity with generic measures, often the activations of all the units in one layer, extracted from ANNs performing the same task or processing the same information. These overall matches between brain activity and ANN hidden states are not adequate to fully reveal the underpinning computational principles guiding each type of system nor do they disentangle the specific contents of the critical computations involved. To investigate human-machine parallels on this algorithmic level, we need to extract the specific contents represented and calculated in ANNs as they simulate a cognitive function, and then determine directly how these relate to the specific neurocomputational structure and dynamics of the human brain executing the same cognitive function.

Language is an especially appropriate domain for such a comparison, both because of its central importance as a biologically unique human capacity, but also because language comprehension involves the dynamic interplay between multifaceted representations of various types of linguistic and nonlinguistic information in a specific context (Kuperberg, 2007; Kuperberg and Jaeger, 2016; Tyler and Marslen-Wilson, 1977), which seems to fit ANNs' competence in flexibly combining different types of features embedded in their rich internal representations (Bengio et al., 2021; Elman, 1990; 1993). Relevant to this, the breakthrough of recent language ANNs seems largely due to their deep contextualized representations of the input which potentially instantiate the underlying linguistic regularities consistent with the specific contents of the input (Lin et al., 2019; Linzen and Baroni, 2020; Manning et al., 2020).

In the research reported here, we seek to implement a set of in-depth comparisons to investigate the algorithmic commonalities between brains and ANNs in terms of building a structured interpretation of a spoken sentence as it unfolds word-by-word over time. A widely accepted "broad church" view in the cognitive sciences, known as the *constraint-based* approach to sentence processing (Altmann, 1998; MacDonald et al., 1994; Trueswell and Tanenhaus, 1994), suggests that the human real-time interpretation of an utterance is subject to multiple types of lexically-driven probabilistic constraints (e.g., syntax, semantics, world knowledge) emerging as the sentence is heard, and where it is the *interpretative coherence* of these constraints that is the basis for successful language comprehension. Depending on the specific context, different sources of constraint will be more or less decisive in selecting from among the grammatically possible structures being considered over time (Altmann and Mirkovic, 2009). Working within this general framework, we investigate the degree of parallelism between brains and ANNs in terms of whether and how the internal operations that in each case implement incremental structural interpretations reflect and integrate the multifaceted probabilistic constraints accumulated in an unfolding sentence.

Specifically, we constructed sentences with varying structures and presented them to both human listeners and to BERT - a deep language ANN (Devlin et al., 2018). For humans, we used source-localized electro-/magnetoencephalography (EMEG) to measure their fleeting brain activity while listening to these sentences and collected their structural interpretations at different positions in a sentence in separate behavioral tests. For BERT, we presented each sentence word-by-word, similarly to how it is received by a human listener. From BERT hidden states, we extracted representations of detailed incremental structure at various positions in a sentence using a structural probe (Hewitt and Manning, 2019), and visualized the dynamic structural interpretation as the structure unfolds via its trajectory in the model space of BERT. We then evaluated the detailed BERT structural measures in terms of the hypothesized *constraint-based* approach, i.e., the strategy adopted by humans to build the structured interpretation of a spoken sentence.

Critically, using Representational Similarity Analysis (RSA) (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008a), we can then directly compare the representational geometry of BERT structural measures with those of spatiotemporally resolved brain activity. In doing so – and in line with previous cross-species comparisons using RSA (Kriegeskorte et al., 2008b) – we assume that significant fits indicate the existence of shared computational contents and styles between BERT and human listeners in building the structure of an unfolding sentence. Therefore, beyond generic correlations between brain activity and ANN hidden states, this combination of methods provides the level of neurocomputational specificity required to elucidate the algorithmic parallels between the human brain and ANNs in terms of a fundamental aspect of incremental language comprehension.

Results

We constructed 60 sets of sentences with varying sentential structures (see Methods) and presented them as spoken sequences to human listeners and word-by-word to BERT. These natural spoken sentences, with their structural interpretations driven by the multifaceted constraints accumulated from the diverse lexical properties that are activated as each spoken word is uniquely identified in the speech [i.e., uniqueness point, UP (Marslen-Wilson, 1987)], provide a realistic simulation of the environment of daily language use. This makes them an ecologically valid basis (Hamilton and Huth, 2020; Nastase et al., 2020) for evaluating the potential alignments between the human brain and ANNs as they represent and interpret the structure of natural language inputs.

In each stimulus set, there are two target sentences differing only in the first verb (Verb1) encountered (Figure 1A). In the HiTrans sentences, Verb1 has high transitivity (e.g., "*found*") and strongly prefers a direct object, while in the LoTrans sentences Verb1 has low transitivity (e.g., "*walked*"). Critically, (a) the structural interpretation of these sentences is potentially 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 probabilistic constraint as more of the sentence is heard. In the example sentences, 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 1B). Because *find* is primarily a transitive verb, the human listener will 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), will 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 ("walking the dog").

This initial structural interpretation of Verb1 does not, however, just depend on linguistic knowledge such as Verb1 transitivity. It also depends on the broader "thematic role" properties of the subject noun – that is, how likely the subject is (or is not) to adopt the Active (agent) role to perform the specified action (Dowty, 1991). This likelihood, of the event structure implied by the different structural combinations of the subject noun and Verb1, will depend on wide ranging "knowledge of the world", linked to the specific words being heard. So, regardless of Verb1 transitivity, the Active interpretation should be more strongly favored in "*The king found/walked…*" given the greater implausibility of a Passive interpretation involving a "*king*" relative to a "*dog*". The word-by-word interpretation of the structure of the sentence – 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 world knowledge and linguistic knowledge.

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 structure being constructed. Thus, for a HiTrans sentence, a Passive interpretation will become more preferred, since a highly transitive Verb1 tends to be interpreted as a passive verb (i.e., the head of a reduced relative clause) when there is the absence of an expected direct object. Conversely, in the LoTrans sentence, the Active interpretation of Verb1 is strengthened by the incoming prepositional phrase, which is fully consistent with the verb's intransitivity and with the event structure conjured up by the sequence of words so far. Hence, these two sentence types are expected to differ in the structural interpretation preferred by the end of the prepositional phrase. However, with the appearance of the actual main verb ("*was covered*" in the example sentences), the preferred Active interpretation of Verb1 as the main verb (as in the LoTrans example) will be rejected, instantiating a Passive interpretation for the LoTrans sentences and confirming existing preferences for the Passive interpretation in the HiTrans sentences.

Human incremental structural interpretations

To validate empirically this account of the evolution of structural interpretations in the constraint-based processing environment, 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, and produced a continuation at each gating point to complete the sentence (see Methods). The listeners' incremental structural interpretations at each point can be inferred from the continuations they provide. For example, the more often a main verb is given in the continuations following the prepositional phrase (e.g., after "*... in the park...*"), the stronger the preference for a Passive interpretation at this point of the sentence. We found that Passive and Active interpretations were generally not all-or-none, varying in plausibility in both HiTrans and LoTrans sentences before the actual main verb is presented (Figure S1). These variations reflect the probabilistic constraints jointly placed by the specific combination of subject noun, Verb1, and the prepositional phrase in each sentence.

To relate these preferences to the broader landscape of distributional language data, we developed corpus-based measures of subject noun thematic role preference and Verb1 transitivity for each sentence, from which we derived a Passive index and an Active index. These indices capture the interpretative coherence between these two types of lexical properties as they affected Passive and Active interpretations separately (see Methods). Both high subject noun agenthood and high Verb1 intransitivity coherently prefer an Active interpretation as the prepositional phrase is heard (i.e., high Active index), and vice versa for the Passive interpretation (i.e., high Passive index). Based on the pre-test results, we found that human interpretations for the two types of sentences as one group were significantly correlated with these quantitative measures of multifaceted constraints, further indicating that human listeners were incrementally interpreting the sentences in the way predicted by the constraint-based approach (Figure S1).

BERT incremental structural interpretations

We then obtained incremental structural interpretations of the same sentences by BERT, aiming at extracting detailed structural measures of each unfolding sentence from the hidden states of BERT. Typically, the structure of a sentence can be represented by a dependency parse tree (De Marneffe et al., 2006) (Figure 1B) in which words are situated at different depths given the structural dependency between them. 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 captures the syntactic dependency relation between each pair of words and abstracts away from the idiosyncratic contents in a sentence that constrain its online structural interpretation. The parse depth is always the same integer for words at the same position in sentences with the same structural interpretation (e.g., "*found*" and "*walked*" in either of the two parse trees in Figure 1B).

However, as shown above, human online structural interpretation involves a variety of probabilistic biases reflecting both linguistic knowledge and broader knowledge of the world evoked by the specific contents in a sentence. To investigate how far BERT also incorporated these broader constraints, we adopted a structural probing approach (Hewitt and Manning, 2019). This aims to reconstruct a sentence's structure by estimating each word's parse depth based on their contextualized representations generated by BERT, which explicitly considers the specific sentential contents (see Methods). Note that BERT is a multi-layer language ANN (24 layers in the version used in this study) which may distribute different aspects of its computational solutions over multiple layers. We trained a structural probing model for each layer, and selected the one with the most accurate structural representation while also including its neighboring layers to cover potentially relevant upstream and downstream information. Following this metric, we used the structural measures obtained from layers 12-16 with layer 14 showing the best performance (see Figure S2 and Methods).

We input each sentence word-by-word to the trained BERT structural probing model, focusing on the incremental structural representation being computed as it progressed from Verb1 to the main verb (see the sequence in Figure 1A). 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 (i.e., Passive interpretation). For each type of target sentence, the BERT parse depth of words at the same position formed a distribution ranging around the corresponding context-free parse depths in either Passive or Active interpretations (Figure S3), providing a word-specific rather than position-specific structural representation. Moreover, the BERT parse depth of earlier words in the sentence was updated with each incoming later word, capturing the incrementality of speech comprehension. These distributed BERT parse depths may indicate the probabilistic structural interpretation since they incorporate the specific contents in a sentence. For example, according to the context-free dependency parse tree (Figure 1B), Verb1 in the Passive interpretation is a passive verb with a parse depth of 2, while in the Active interpretation Verb1 is the main verb with a parse depth of 0. Therefore, a BERT parse depth that puts Verb1 closer to 2 suggests that a Passive interpretation is preferred, while a Verb1 parse depth closer to 0 indicates preference for an Active interpretation.

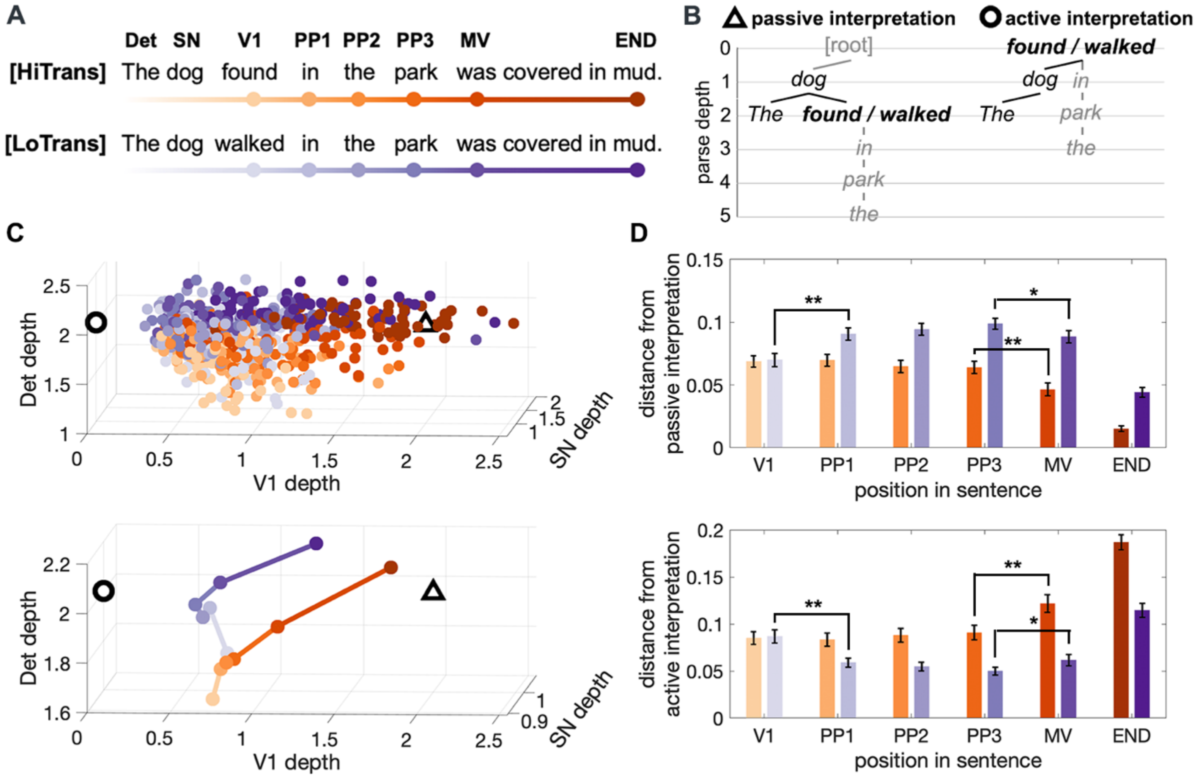

Figure 1. Incremental interpretation of sentential structure by BERT. (A) An example set of target sentences differing in the transitivity of Verb1 (V1), HiTrans: high V1 transitivity, LoTrans: low V1 transitivity. Lightness of dot encodes different positions in the sentence. Det: determiner, SN: subject noun, PP1-PP3: prepositional phrase, MV: main verb, END: the last word in the sentence. **(B)** 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. **(C)** Incremental interpretations of the dependency between SN and V1 in the model space consisting of the parse depth of Det, SN and V1. Upper: Each colored small circle represents the parse depth vector up to V1 derived at a certain position in the sentence [with the same color scheme as in (A)]. The triangle and circle represent the context-free dependency parse vectors for Passive and Active interpretations in (B). Lower: incremental interpretations of the two types of target sentences represented by the trajectories of median parse depth. **(D)** 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 (C)] (two-tailed two-sample t-test, *: P < 0.05, **: P < 0.001, error bars represent SEM).

We further quantified and visualized BERT's word-by-word structural interpretations, focusing on the dependency between the subject noun and Verb1 which is core to the structural interpretation here – whether the subject noun is the agent or the patient of Verb1. We represented each sentence by a 3-dimensional vector including the BERT parse depth of the first three words up to Verb1. This 3D vector was kept updated with every incoming word, capturing the degree to which the structural dependency between the subject noun and Verb1 was dynamically interpreted (and potentially re-interpreted) given the contents of the subsequent words in the sentence. Crucially, the trajectory of each sentence in this 3D model space, as it unfolds word-by-word, characterizes the dynamic structural interpretation constructed by BERT (Figure 1C, upper). We found considerably intertwined trajectories of individual HiTrans and LoTrans sentences, suggesting that BERT structural measures are sensitive to the specific 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 separately and located them in the model space as landmarks (Figure 1C), so that the plausibility of either interpretation for an unfolding sentence can be estimated by its distance from the corresponding landmark. As shown by the trajectories of the median BERT parse depth of the two sentence types (Figure 1C, lower), HiTrans sentences moved unidirectionally towards the Passive interpretation landmark after Verb1, with significant changes of distances found at the main verb (Figure 1D, orange bars). 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 found at both Verb1 and main verb (Figure 1D, purple bars). These results resemble the pattern of human interpretative preference observed in the continuation pre-tests, where the Passive and Active interpretations are separately preferred in HiTrans and LoTrans sentences by the end of the prepositional phrase in a probabilistic manner (Figure S1), and the structural ambiguity is eventually resolved with the appearance of the actual main verb.

Going beyond their behavioral resemblance to human structural interpretations of the same sentences, we investigated whether BERT structural interpretations could also be directly explained by the multifaceted constraints placed by the subject noun and Verb1 (see Methods). We first focused on BERT's interpretative mismatch quantified by a sentence's distance from each of the two landmarks in the model space, which was dynamically updated as the sentence unfolded (Figure 1C). Consistently, from the incoming prepositional phrase to the main verb, the unfolding sentences that are closer to the Passive landmark have higher Verb1 transitivity, higher Passive index and lower Active index, while those closer to the Active interpretation landmark tend in the opposite direction (Figure 2A). Moreover, the change of distance towards one interpretation landmark between two consecutive words is also correlated with these constraints (Figure 2C and 2D) (see Figure S4 for results of all BERT layers). Significant effects were primarily seen when the first word of the prepositional phrase is encountered, suggesting that this leads to an immediate update in the structural interpretation, in combination with the accumulated constraints from the preceding subject noun and Verb1.

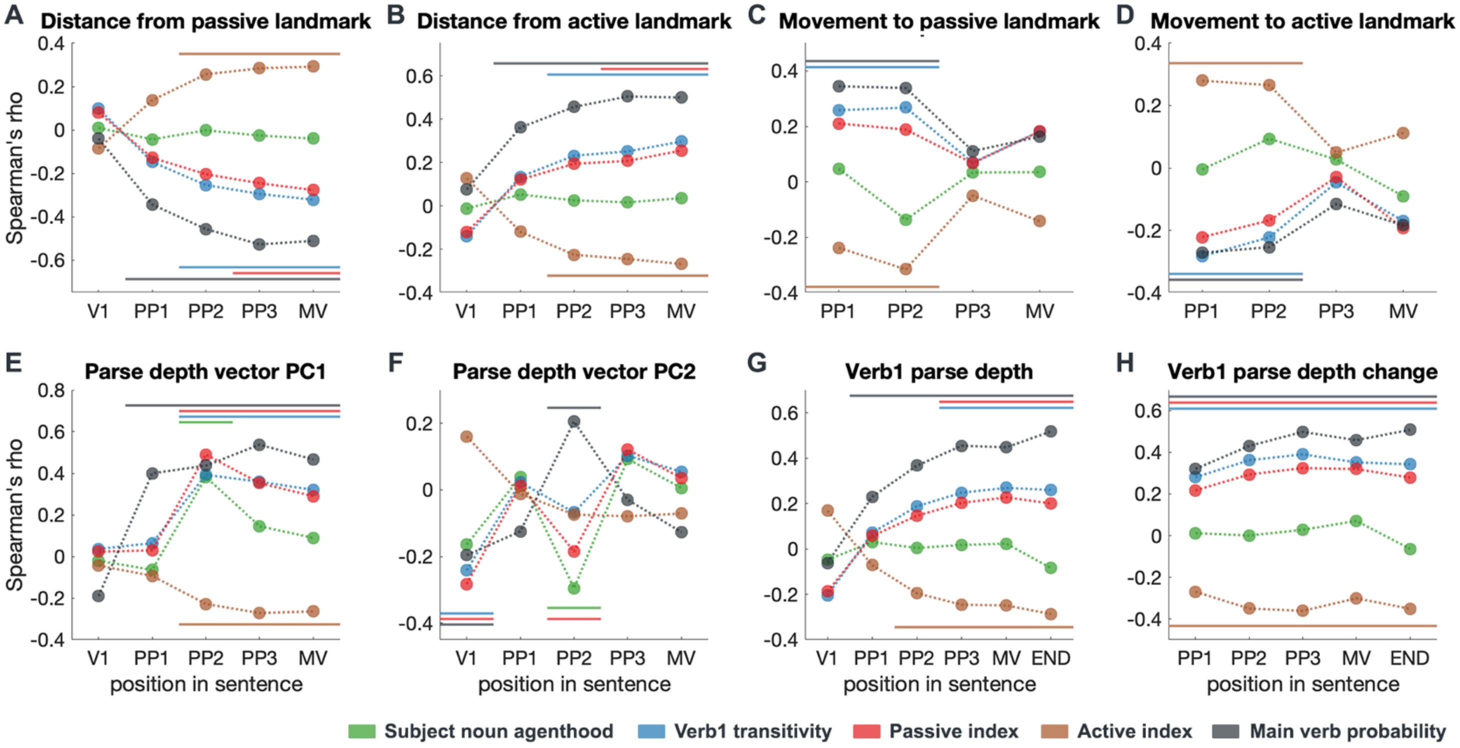

Figure 2. Correlation between incremental BERT structural measures and explanatory variables. BERT structural measures include **(A, B)** BERT interpretative mismatch represented by each sentence's distance from the two landmarks in model space (Figure 1C); **(C, D)** Dynamic updates of BERT interpretative mismatch represented by each sentence's movement to the two landmarks; **(E, F)** Overall structural representations captured by the first two principal components of BERT parse depth vectors; **(G, H)** BERT V1 parse depth and its dynamic updates. Explanatory variables include lexical constraints derived from massive corpora and the main verb probability derived from human continuation pre-tests (Spearman correlation, permutation test, $P_{FDR} < 0.05$, multiple comparisons corrected for all BERT layers, results shown here are based on layer 14, see Figures S4-S6 for the results of all layers).

Similarly, we also found that both the incremental BERT parse depth vectors as a whole (which are captured by their principal components) and the contextualized parse depth of Verb1 (which is the most indicative marker of BERT structural interpretation) are correlated with the constraints of subject noun and Verb1 (Figure 2E-H) (See Figures. S5 and S6 for results of all BERT layers). Moreover, the significant effects consistently found as the sentence unfolds suggest that properties of preceding words are carried over and used to constrain the interpretation of the upcoming input, which is key to resolving potential discontinuous structural dependencies. In addition, we found that main verb probability in the human continuations after the prepositional phrase, which directly reflects the probabilistic interpretations of human listeners, also matched to the structural interpretations generated by BERT (black bars in Figure 2).

Taken together, BERT structural interpretations, as well as how they are developed and updated as the sentence unfolds word-by-word, can be explained by the same multifaceted probabilistic constraints that constrain incremental structural interpretations in human listeners, suggesting that related forms of constraint-based approaches could be adopted by both humans and BERT to build structured interpretations.

Neural alignment for incremental interpretations in BERT and human listeners

Beyond the behavioral alignment between structural interpretations by humans and BERT, we carried out RSA to test the detailed BERT structural measures against source-localized EMEG recordings collected when the same sentences were incrementally delivered to human listeners. Specifically, we compared the representational geometry of BERT structural measures and the multifaceted constraints they incorporate with that of spatiotemporally resolved brain activity. Significant RSA fits reveal where and when in the brain there is an alignment between the structural interpretations by BERT and humans, suggesting the existence of common underlying computational strategies. Compared with the overall activations from the hidden layers, the detailed BERT structural measures provide better neurocomputational specificity for investigating the parallels between ANNs and brains in building structured interpretations.

As shown above, consistent with the *constraint-based* approach, the context-specific BERT parse depths incorporate both linguistic knowledge and broad world knowledge that constrain the incremental sentential structure underpinning the ongoing event interpretation. Thus, looking beyond the neural dynamics of the core fronto-temporal core language network, RSA based on these broader BERT structural measures enables us to address how brain regions encoding non-linguistic knowledge (e.g., event-related knowledge and episodic memory) contribute to building a structured interpretation over time. Given the probabilistic interpretations by BERT and human listeners reported above, we combined HiTrans and LoTrans sentences as one group in the following analyses rather than treating them as two categories.

To test these BERT structural measures in the brain, we began with the BERT parse depth vector containing the parse depth of all the words in an incremental input, providing a structural representation for the input delivered so far. Second, we tested the interpretative mismatch quantified by the cosine distance between an incremental BERT parse depth vector and the corresponding incremental context-free parse depth vector for the Passive and the Active interpretations. The degree of such mismatch is proportional to the broader evidence for or against the structural interpretation being incrementally considered (the smaller the distance, the more positively loaded the corresponding interpretation). We also particularly focused on the BERT parse depth of Verb1. This is updated with each incoming word, with increased or decreased depth reflecting the preference biased to a Passive or an Active interpretation separately (Figure S7).

For human neural data, we focused on source-localized EMEG recorded at three critical positions in each sentence: (a) Verb1 – when its structural dependency with the preceding subject noun was initially established despite potential ambiguity, (b) the preposition (introducing the prepositional phrase following Verb1) – 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 structural ambiguity was definitively resolved with the identification of the intended structure. We aligned the continuous EMEG data to the onset of Verb1, the preposition and the main verb respectively and obtained three epochs each 600ms in length.

As revealed by RSA, the incremental BERT parse depth vectors exhibited significant fits to brain activity consistently in all three epochs (Figure 3A-C). In the Verb1 epoch, effects in bilateral frontal and temporal regions started immediately from Verb1 onset and continued until the uniqueness point, while the parse depth of Verb1 per se showed similar but more sustained effects which peaked at Verb1 uniqueness point when the word had been identified (Figure S8). Whereas effects in the preposition and main verb epochs were found in the left hemisphere after the uniqueness point.

Turning to the interpretative mismatch for the two possible interpretations, we only observed significant effects of the mismatch for the Active interpretation in the Verb1 epoch (Figure 3D). However, it was the mismatch for the Passive interpretation that fitted brain activity in the preposition and main verb epochs (Figure 3E and 3F, marginal significance in main verb epoch with cluster-wise P = 0.06). These results suggest that listeners tend to have an initial preference for an Active interpretation but may start favoring a Passive interpretation when the prepositional phrase begins to be heard, which is consistent with the tendency to process the first noun or encountered in a sentence as the subject or agent (Bever, 1970; Jackendoff and Jackendoff, 2002; Karmiloff-Smith, 1981).

Effects of the BERT parse depth vectors and those of the mismatch for the preferred structural interpretation have substantial overlaps in terms of their spatio-temporal patterns in the brain, characterized primarily by a transition from bilateral to left-lateralized fronto-temporal regions. 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 presumably resolved structure also involved regions in the left prefrontal and inferior parietal cortex (Figure 3C) which are involved in the multiple-demand network (Duncan, 2010).

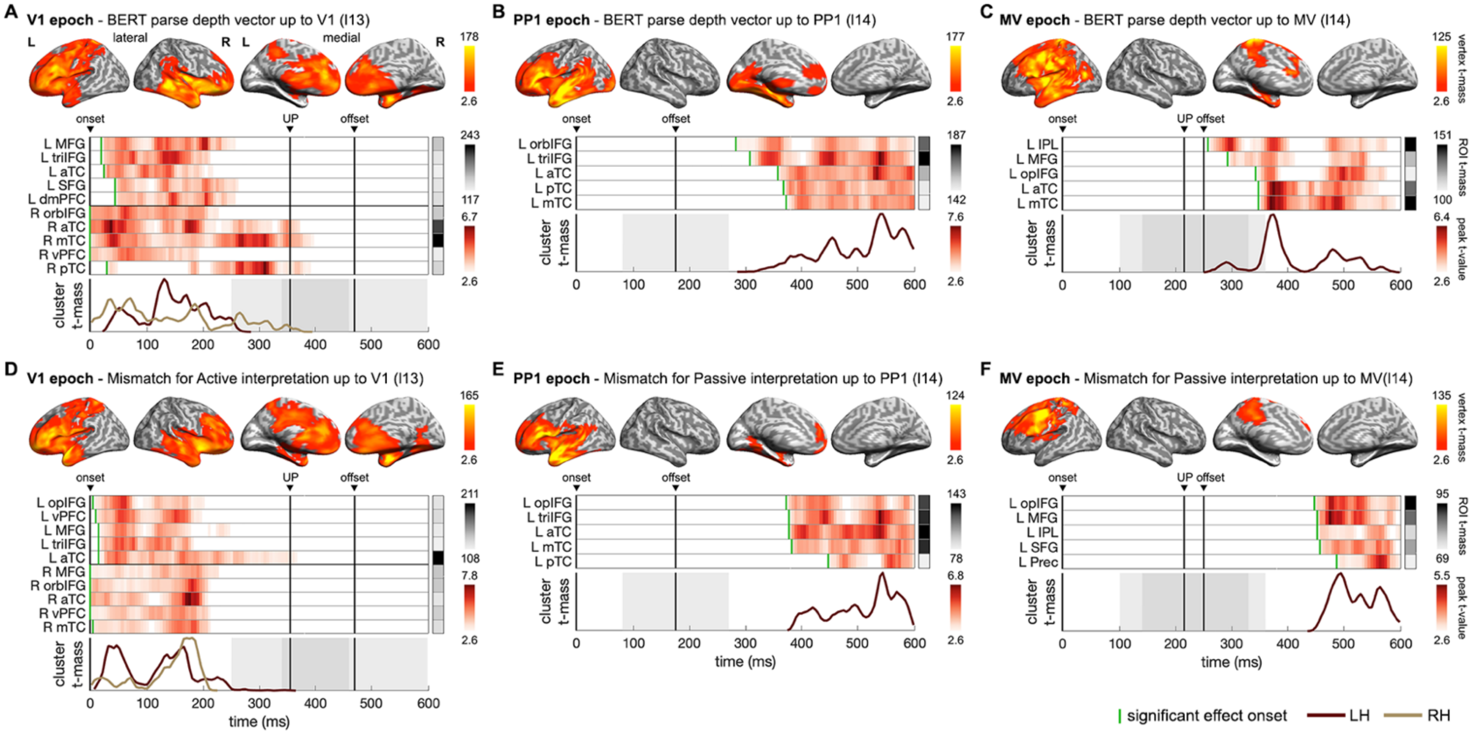

Figure 3. Neural dynamics underpinning the emerging structure and interpretation of an unfolding sentence. (A-C) ssRSA results of BERT parse depth vector up to V1, PP1 and MV in epochs separately time-locked to their onsets. **(D-F)** ssRSA results of the mismatch for the preferred structural interpretation (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); marginally significance in (F) with cluster-wise $P = 0.06$]. Solid vertical lines indicate the timings of onset, mean uniqueness point (UP) and offset of the word time-locked in the epoch with grey shades indicating the range of one SD. LH/RH: left/right hemisphere. See Table S1 for full anatomical labels. See Figure S9 for the significant results of other BERT layers in the MV epoch.

In fact, the potential ambiguity between a Passive and an Active interpretation is a matter of whether Verb1 is considered as a passive verb or the "main verb". This is resolved when the actual main verb is encountered. We probed how this is achieved in the brain using the dynamic BERT parse depth of Verb1. Specifically, the cognitive demands required by the resolution process can be characterized by the change between the updated BERT parse depth of Verb1 when the actual main verb is heard and its initial value when the sentence just proceeds to Verb1. Therefore, we 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 main verb uniqueness point, and then extended to more anterior temporal regions (Figure 4A). Following the main verb offset, in the left anterior temporal region, the declining effects of the change of Verb1 parse depth seamlessly overlapped with the increasing effects of the updated Verb1 parse depth (Figure 4B and 4C). It is also worth noting that the left hippocampus was activated for both the change of Verb1 parse depth and the updated parse depth of Verb1 after the actual main verb is presented, suggesting that episodic memory of experienced events might contribute to updating the structured interpretation of a sentence (Altmann and Ekves, 2019; Bicknell et al., 2010; Metusalem et al., 2012).

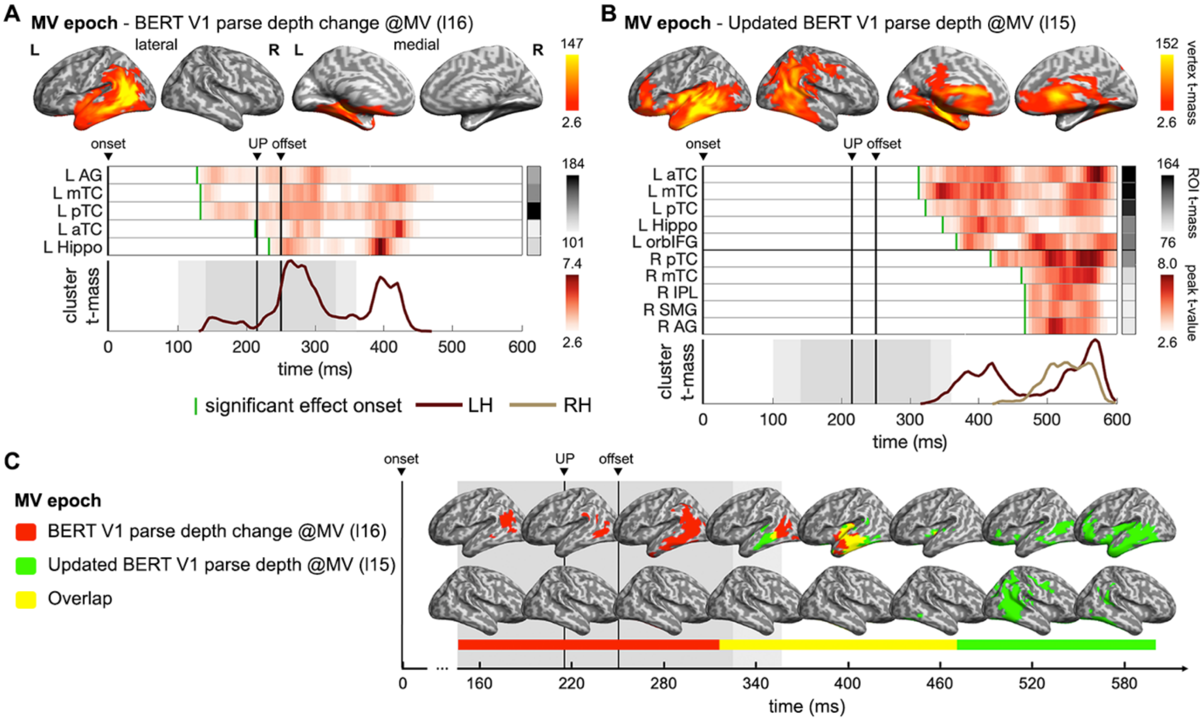

Figure 4. Neural dynamics updating the incremental structural interpretation. (A) ssRSA results of BERT V1 parse depth change at MV (difference between the BERT parse depth of V1 when the sentence input unfolded to V1 and to MV). **(B)** ssRSA results of the updated BERT V1 parse depth when the input sentence unfolded to MV. **(C)** Spatio-temporal overlap between the effects in (A) and (B) (cluster-based permutation test, vertex-wise P < 0.01, cluster-wise P < 0.05).

Emergent structural interpretations driven by multifaceted constraints in the brain

By leveraging detailed BERT structural measures, we revealed the neural dynamics underlying the incremental structural interpretation of an unfolding sentence. Following this up, we asked how the multifaceted constraints incorporated in BERT structural measures drive the structured interpretation in human listeners? When and where do these constraints emerge in the brain? How are their neural effects related to those of the resolved sentential structure? Addressing these questions will provide insights into the commonalities in the use of multi-dimensional probabilistic constraints to obtain structured interpretations in both BERT and human listeners.

To this end, we first tested the subject noun thematic role properties. Significant effects of agenthood and patienthood were seen in the preposition epoch (Figure 5A) and in the main verb epoch (Figure 5B) separately. These effects of the subject noun *per se* preceded the effects of BERT parse depth vectors up to the word time-locked in each epoch (compare Figure 5A with Figure 3B, compare Figure 5B with Figure 3C). This indicates that subject noun thematic role was evaluated before building the overall structural interpretation of the incremental input delivered so far. Specifically, the initial preference for an Active interpretation during Verb1, while present as the prepositional phrase started, was superseded by the preference for a Passive interpretation as the rest of the phrase and the main verb 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 actual main verb could differ between sentences (e.g., in "*The dog found in the park...*" and "*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 quantifies the degree of interpretative coherence for the preferred interpretation, whether Passive or Active (see Methods). Thus, a higher value only indicates greater interpretative coherence between the subject noun and Verb1 regardless of which interpretation is involved.

Effects of this non-directional interpretative coherence measure appeared very soon after the main verb onset in both hemispheres and lasted till main verb offset (Figure 5C), suggesting an immediate re-evaluation of the previously integrated constraints from the subject noun and Verb1 after a listener realized that the sentence had not yet finished. Moreover, these effects roughly co-occurred with the effects of subject noun patienthood (Figure 5B and 5C), indicating that a patient role of 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 is consistent with recent claims that the DMN integrates external information with internal prior knowledge to make sense of an external input 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 from episodic memory (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 5D), suggesting that the intended Passive interpretation was eventually confirmed. Previous studies have revealed that sentence-specific expectancy and general event-relevant inference are processed in the left and right hemispheres separately (Jung-Beeman, 2005; Metusalem et al., 2016). 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 simultaneously as the main verb was recognized (compare Figure 3C with Figure 5D). Therefore, a critical question is whether, and if so, how the online structural interpretation of a specific sentence is facilitated by the interpretative coherence conjured up from broad world knowledge.

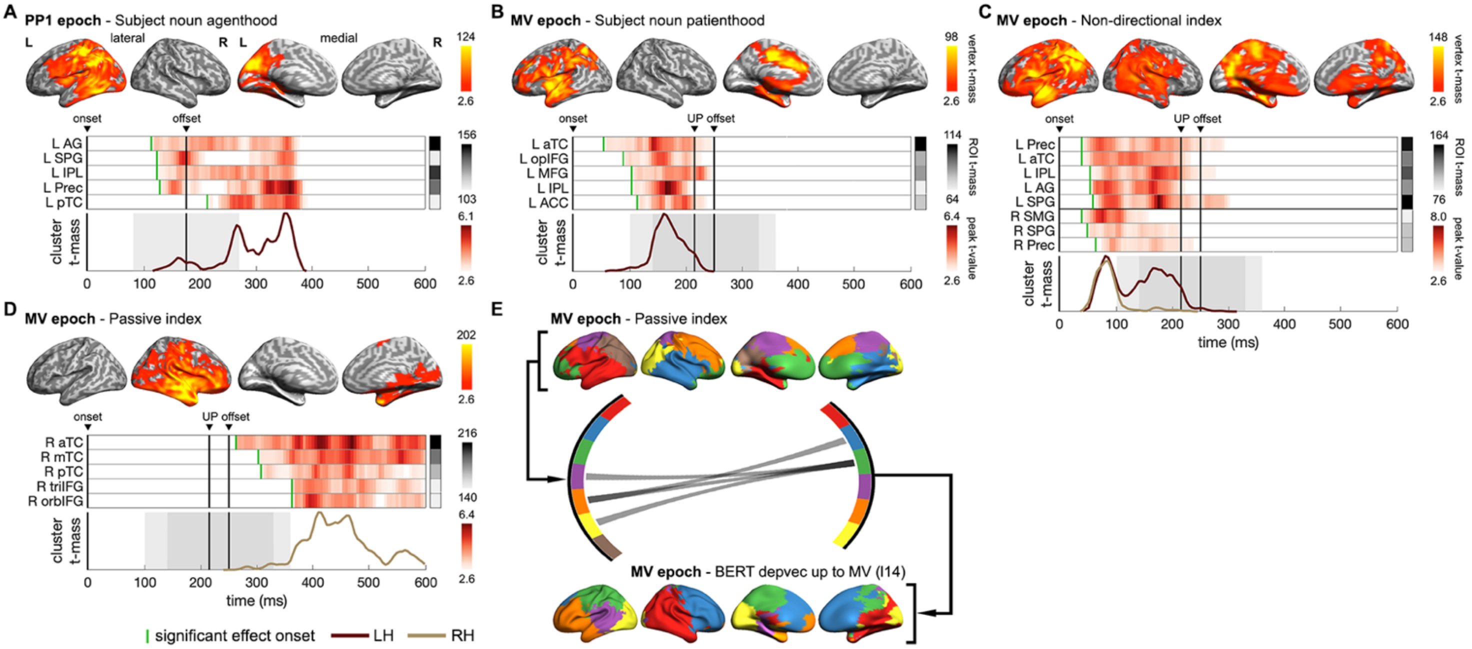

Figure 5. Neural dynamics of multifaceted probabilistic constraints underpinning incremental structural interpretations. (A, B) ssRSA results of SN agenthood and SN patienthood (i.e., plausibility of SN being the agent or the patient of V1) in PP1 and MV epochs separately. **(C)** ssRSA results of non-directional 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 Figure S10 for more details). [(A-D) cluster-based permutation test, vertex-wise $P < 0.01$, cluster-wise $P < 0.05$; (E) permutation test $P_{FDR} < 0.05$]

To address this question, we adopted non-negative matrix factorization to decompose the whole-brain RSA fits of Passive index and BERT parse depth vector in the main verb epoch into two sets of components given their temporal synchronizations (see Methods). We then conducted 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 5E). 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 the construction of sentential structures in the left hemisphere is integrated with the general event knowledge drawn from the right hemisphere (see Figure S10 for more details).

In summary, we found that BERT structural measures exhibited significant fits to brain activity during the incremental development of a structured interpretation in human listeners. Beyond this, and consistent with the *constraint-based* approach, these results also revealed how the multifaceted constraints incorporated in BERT structural measures are activated, integrated and evaluated to constrain the word-by-word interpretation of a sentence's structure in an extensive set of brain regions beyond the classical core fronto-temporal system.

Discussion

In this study, we revealed a deep alignment between human listeners and language ANNs in building structured interpretations from incrementally delivered natural sentences. Both their behaviors and internal structural representations conform to the hypothesized *constraint-based* approach, suggesting shared underlying computational styles. By using detailed structural measures derived from ANN hidden layers rather than their overall activations, we conducted a precise mapping between humans and machines with improved neurocomputational specificity, providing direct empirical evidence about how the internal operations of ANNs do or do not relate to processes in the human brain as listeners understand spoken language on the algorithmic level (i.e., shared computational styles).

Human speech comprehension involves a complex set of processes that transform an auditory input into the speaker's intended meaning, during which each word is sequentially interpreted and integrated with the preceding words to obtain a coherent interpretation (Choi et al., 2021; Lyu et al., 2019). However, rather than simple linear concatenations, individual words are combined according to the nonlinear and often discontinuous structure embedded in a spoken utterance as it is delivered over time (Everaert et al., 2015). 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 (Law and Pylkkanen, 2021; Nelson et al., 2017), manipulating the syntactic complexity in sentences (Pallier et al., 2011), or studying artificial grammatical rules without intelligible contents (Friederici et al., 2006; Makuuchi et al., 2009). But as this and other work have shown, finding the structure in an unfolding sentence also depends on the constraints jointly placed by other linguistic properties of the input and broad world knowledge (Bever, 1970; MacDonald et al., 1994; Tanenhaus et al., 1995; Trueswell and Tanenhaus, 1994; Tyler and Marslen-Wilson, 1977).

This means that to reveal how a spoken sentence is incrementally built and represented in the brain, we need to focus on natural speech in its full ecological complexity. This raises the challenge of developing quantitative representations of sentential structure entailed by the concurrent effects of multifaceted constraints. As a potential solution, we extracted detailed incremental structural measures driven by the contents in each sentence from BERT hidden states. Although BERT is not specifically trained to parse sentences, its successful performance in various NLP tasks suggests that it can acquire from its massive training corpora the multi-dimensional properties (Manning et al., 2020) related to determining the structural dependency among a sequence of input words. We showed directly that BERT structural measures incorporated these constraints and that they matched with both the behaviors and the brain activity of human listeners while listening to the same set of sentences.

In a broad sense, using ANNs to help us understand the neural basis of human cognition complements the long-time pursuit of generative rules and interpretable models in cognitive neuroscience. ANNs have informed in many ways the internal workings of various cognitive functions in the brain by providing quantitative representations and falsifiable predictions that aim to connect cognitive behaviors and relevant neural activity in the brain (Kietzmann et al., 2019; Kriegeskorte and Douglas, 2018; Yang et al., 2019). This seems essential if we are to quantify the outcome of intertwined multi-dimensional regularities in a specific instance (e.g., a spoken sentence) and to construct the representational geometry to be probed in the brain. Where language ANNs are concerned, it is feasible to obtain a dynamic representation of an unfolding input that is updated with each incoming word, such as the BERT parse depth vector. This captures well the incrementality at the core of speech comprehension, that each word is interpreted within the sentential structure built so far and in turn updates the structure being constructed.

Here, we presented empirical evidence suggesting that the structured interpretations by both human listeners and BERT conform to the *constraint-based* approach. But the means by which they acquire and implement the potentially shared computational styles could nonetheless be radically different. According to the recent “direct fit” view (Hasson et al., 2020), both the human brain and ANNs may learn task-relevant structures through dense sampling, which enables humans and machines to flexibly produce appropriate context-dependent behaviors through interpolation within the mental or model space spanned by instances in life experience or training datasets. However, as an immensely complex biological object with a long evolutionary history, the human brain is not a giant undifferentiated network that learns from experience in essentially the same way as ANNs, especially where these are operating in ways that are biologically implausible (Lillicrap et al., 2020). The evolved neurobiological structure in the brain underpins not only task-relevant structures but also rules of abstraction and generalization to the unseen (Mansouri et al., 2020; Pulvermuller et al., 2021).

Following this evolutionary perspective, the Dual Neurobiological Systems framework (Marslen-Wilson and Bozic, 2018) supposes that human language has emerged as an evolutionary coalition between a bi-hemispheric system supporting general social communication based on multi-modal input that is inherited from our primate ancestors and a left hemisphere system underlying the more complex combinatorial capacities (Hauser et al., 2002) that makes human language distinct from its evolutionary precursors. Consistent with this view, we found bi-hemispheric or right-hemisphere dominant effects for computations related to broad world knowledge. Meanwhile, effects of the incremental sentential structure were left-lateralized beyond Verb1 when more complex syntax rather than canonical linear adjacency is required to build a structured interpretation. Therefore, underneath the similar structural representations on the algorithmic level (Marr, 1982), the actual implementation of the *constraint-based* approach in the brain is subject to the neurobiology of language, rather than being arbitrarily folded into billions of parameters in ANNs. Reconciling this disparity may be central to future attempts to improve AI systems for seamless communication with humans.

In conclusion, through this uniquely in-depth investigation of the parallels between the brain and ANNs, we revealed commonalities in their structured interpretations driven by multi-dimensional probabilistic constraints. In doing so, we demonstrated the value of using task-specific representations in ANNs to address neural dynamics underpinning higher-level cognitive functions with improved neurocomputational specificity, providing realistic neurobiological constructs for the development of future AI systems.

Materials and Methods

Details of materials and methods are provided in Supplementary Information.

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). 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 and Nai Ding for helpful discussions.

Author contributions

Conceptualization: L.K.T., W.D.M., B.L.
Investigation, Data curation: B.L., Y.F.
Methodology, Software, 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.

References

Altmann, G.T.M. (1998). Ambiguity in sentence processing. Trends Cogn Sci *2*, 146-152.

Altmann, G.T.M., and Ekves, Z. (2019). Events as intersecting object histories: A new theory of event representation. Psychol Rev *126*, 817-840.

Altmann, G.T.M., and Mirkovic, J. (2009). Incrementality and Prediction in Human Sentence Processing. Cognitive Sci *33*, 583-609.

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

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

Bever, T.G. (1970). The cognitive basis for linguistic structures. In Cognition and the Development of Language, J.R. Hayes, ed. (John Wiley).

Bicknell, K., Elman, J.L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. J Mem Lang *63*, 489-505.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. ArXiv.

Caucheteux, C., and King, J.-R. (2020). Language processing in brains and deep neural networks: computational convergence and its limits. bioRxiv.

Choi, H.S., Marslen-Wilson, W.D., Lyu, B., Randall, B., and Tyler, L.K. (2021). Decoding the Real-Time Neurobiological Properties of Incremental Semantic Interpretation. Cereb Cortex *31*, 233-247.

Cichy, R.M., and Kaiser, D. (2019). Deep Neural Networks as Scientific Models. Trends Cogn Sci *23*, 305-317.

De Marneffe, M.-C., MacCartney, B., and Manning, C.D. (2006). Generating typed dependency parses from phrase structure parses. held in Genoa, Italy, (European Language Resources Association), pp. 449-454.

Devereux, B.J., Clarke, A., and Tyler, L.K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Sci Rep *8*, 10636.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv.

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 Cogn Sci *14*, 172-179.

Elman, J.L. (1990). Finding Structure in Time. Cognitive Sci *14*, 179-211.

Elman, J.L. (1993). Learning and development in neural networks: the importance of starting small. Cognition *48*, 71-99.

Everaert, M.B.H., Huybregts, M.A.C., Chomsky, N., Berwick, R.C., and Bolhuis, J.J. (2015). Structures, Not Strings: Linguistics as Part of the Cognitive Sciences. Trends Cogn Sci *19*, 729-743.

Friederici, A.D., Bahlmann, J., Heim, S., Schubotz, R.I., and Anwander, A. (2006). The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proc Natl Acad Sci U S A *103*, 2458-2463.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2021). Thinking ahead: prediction in context as a keystone of language in humans and machines. bioRxiv.

Guclu, U., and van Gerven, M.A. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J Neurosci *35*, 10005-10014.

Hamilton, L.S., and Huth, A.G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn Neurosci *35*, 573-582.

Hasson, U., Nastase, S.A., and Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron *105*, 416-434.

Hauser, M., Chomsky, N., and Fitch, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? Science *298*, 1569-1579.

Hewitt, J., and Manning, C.D. (2019). A Structural Probe for Finding Syntax in Word Representations. held in Minneapolis, Minnesota, (Association for Computational Linguistics), pp. 4129-4138.

Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process Mag *29*, 82-97.

Humphreys, G.F., Lambon Ralph, M.A., and Simons, J.S. (2021). A Unifying Account of Angular Gyrus Contributions to Episodic and Semantic Cognition. Trends Neurosci *44*, 452-463.

Jackendoff, R., and Jackendoff, R.S. (2002). Foundations of language: Brain, meaning, grammar, evolution (Oxford University Press, USA).

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. Trends Cogn Sci *9*, 512-518.

Karmiloff-Smith, A. (1981). A functional approach to child language: A study of determiners and reference (Cambridge University Press).

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol *10*, e1003915.

Kietzmann, T.C., Spoerer, C.J., Sorensen, L.K.A., Cichy, R.M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proc Natl Acad Sci U S A *116*, 21854-21863.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annu Rev Vis Sci *1*, 417-446.

Kriegeskorte, N., and Douglas, P.K. (2018). Cognitive computational neuroscience. Nat Neurosci *21*, 1148-1160.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci *17*, 401-412.

Kriegeskorte, N., Mur, M., and Bandettini, P.A. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci *2*, 4.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126-1141.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM *60*, 84-90.

Kuperberg, G.R. (2007). Neural mechanisms of language comprehension: challenges to syntax. Brain Res *1146*, 23-49.

Kuperberg, G.R., and Jaeger, T. (2016). What do we mean by prediction in language comprehension? Lang Cogn Neurosci *31*, 32-59.

Law, R., and Pylkkanen, L. (2021). Lists with and without Syntax: A New Approach to Measuring the Neural Processing of Syntax. J Neurosci *41*, 2186-2196.

Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., and Hinton, G. (2020). Backpropagation and the brain. Nat Rev Neurosci *21*, 335-346.

Lin, Y., Tan, Y.C., and Frank, R. (2019). Open Sesame: Getting inside BERT's Linguistic Knowledge. held in Florence, Italy, (Association for Computational Linguistics), pp. 241-253.

Linzen, T., and Baroni, M. (2020). Syntactic Structure from Deep Learning. Annu Rev Linguist *7*, 195-212.

Lyu, B., Choi, H.S., Marslen-Wilson, W.D., Clarke, A., Randall, B., and Tyler, L.K. (2019). Neural dynamics of semantic composition. Proc Natl Acad Sci U S A *116*, 21318-21327.

MacDonald, M.C., Pearlmutter, N.J., and Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. Psychol Rev *101*, 676-703.

Makuuchi, M., Bahlmann, J., Anwander, A., and Friederici, A.D. (2009). Segregating the core computational faculty of human language from working memory. Proc Natl Acad Sci U S A *106*, 8362-8367.

Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. Proc Natl Acad Sci U S A *117*, 30046-30054.

Mansouri, F.A., Freedman, D.J., and Buckley, M.J. (2020). Emergence of abstract rules in the primate brain. Nat Rev Neurosci *21*, 595-610.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. Cognition *25*, 71-102.

Marslen-Wilson, W.D., and Bozic, M. (2018). Dual neurobiological systems underlying language evolution: inferring the ancestral state. Curr Opin Behav Sci *21*, 176-181.

Matchin, W., and Hickok, G. (2020). The Cortical Organization of Syntax. Cereb Cortex *30*, 1481-1498.

Metusalem, R., Kutas, M., Urbach, T.P., and Elman, J.L. (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, T.P., Hare, M., McRae, K., and Elman, J.L. (2012). Generalized event knowledge activation during online sentence comprehension. J Mem Lang *66*, 545-567.

Nastase, S.A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. Neuroimage *222*, 117254.

Nelson, M.J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C., and Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. Proc Natl Acad Sci U S A *114*, E3669-E3678.

Pallier, C., Devauchelle, A.D., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci U S A *108*, 2522-2527.

Pulvermuller, F., Tomasello, R., Henningsen-Schomers, M.R., and Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. Nat Rev Neurosci *22*, 488-502.

Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat Neurosci *22*, 1761-1770.

Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? Nat Rev Neurosci *22*, 55-67.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. bioRxiv.

Schrimpf, M., Kubilius, J., Lee, M.J., Ratan Murty, N.A., Ajemian, R., and DiCarlo, J.J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. Neuron *108*, 413-423.

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

Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR *abs/1409.1556*.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., and Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science *268*, 1632-1634.

Trueswell, J.C., and Tanenhaus, M.K. (1994). Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. In Perspectives on sentence processing., (Lawrence Erlbaum Associates, Inc), pp. 155-179.

Tyler, L.K., and Marslen-Wilson, W.D. (1977). The On-Line Effects of Semantic Context on Syntactic Processing. J Verbal Learn Verbal Behav *16*, 683-692.

Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci *19*, 356-365.

Yang, G.R., Joglekar, M.R., Song, H.F., Newsome, W.T., and Wang, X.J. (2019). Task representations in neural networks trained to perform many cognitive tasks. Nat Neurosci *22*, 297-306.

Yang, G.R., and Wang, X.J. (2020). Artificial Neural Networks for Neuroscientists: A Primer. Neuron *107*, 1048-1070.

Yeshurun, Y., Nguyen, M., and Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. Nat Rev Neurosci *22*, 181-192.