Modeling pronoun resolution in the brain

Jixing Li
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates

Wen-Ming Luh
National Institute on Aging
Baltimore, MD, USA

Liina Pylkkänen
New York University
New York, USA
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates

Yiming Yang*
Jiangsu Normal University
Xuzhou, Jiangsu, China

John Hale*
University of Georgia
Athens, GA, USA

Abstract

Our ability to ascertain which person a pronoun refers to is a central part of human language understanding. Toward a process-based understanding of the brain's pronoun-resolution abilities, we evaluated four computational models against brain activity during naturalistic comprehension. These models each formalizes a different strand of explanation for pronoun resolution that has figured in the cognitive and linguistic literature. These include syntactic binding constraints, discourse coherence and principles of memory retrieval. We also examined a deep neural network model that has shown high performance in Natural Language Processing. We collected both functional Magnetic Resonance Imaging (fMRI) and magnetoencephalography (MEG) data while English and Chinese speakers listened to an extended narrative in the scanner. We applied univariate and multivariate analyses to correlate model predictions with brain activity patterns time-locked at each third person pronoun in the narratives. Our combined results all favor the memory-based model, suggesting a domain-general mechanism for pronoun resolution that resembles memory retrieval.

Introduction

One unique machinery of human language is reference, that is, using a linguistic symbol such as pronouns to pick out certain entities in the discourse context. Pronouns cannot be interpreted by themselves and depend their meanings on an antecedent expression. We typically have no difficulty linking a pronoun to its antecedent during language comprehension, yet the neural computations underlying this linking process remain elusive. To achieve a detailed, process-based understanding of pronoun resolution, we utilize computational models that lay out specified and carefully thought-out steps to achieve pronoun resolution. By evaluating the cognitive validity of these models against human brain activity, we provide insights on the constituting elements and their interactions during the cognitive process of pronoun resolution.

We selected three symbolic models each formalizing a different strand of explanation for pronoun resolution that has figured in the cognitive and linguistic literature. The syntax-based Hobbs model[1] implements the classic Binding Theory[2] in formal linguistics, which states that pronouns cannot be coindexed with antecedents in the same clause. For example, in "Mary loves her", the pronoun "her" cannot refer back to the "Mary". The discourse-based Centering model[3] implements the Centering Theory[4] that views pronominalization as a means to achieve discourse coherence, such that the most prominent entity is maintained through the use of pronouns in connected sentences. The memory-based ACT-R model[5] conforms to the salience account for pronoun resolution, and selects the most highly-activated entity in the working memory as the antecedent of the pronoun. In addition to the knowledge-based models, we also included one data-oriented deep neural network model that has shown high performance in Natural Language Processing. The neural network model[6,7] (henceforth the NeuralCoref model) considers all spans of words in a document as possible mentions and learns a distribution over possible antecedents for each mention in a labeled dataset.

We correlated the model predictions with brain activities during pronoun resolution. We recorded the blood-oxygen-level-dependent (BOLD) signals while participants listened to a 100-minute audiobook of "The Little Prince" in the fMRI scanner. This naturalistic setting is ideal for comparing computational models for pronoun resolution in an extended narrative. We collected data from both English and Chinese speakers using the exactly same paradigm to further test whether linguistic typology would influence the strategies for pronoun resolution. To explore the temporal dynamics of the model fit, we also collected an MEG dataset while

English speakers listened to a 12-minute audio excerpt from the YouTube channel "SciShow Kids". We applied both multivariate representational similarity analyses (RSA)[8] and univariate general linear model (GLM) analyses to compare the four models' relatedness to the BOLD responses and the source-localized MEG data time-locked at each third person pronoun in the narratives (see Figure 1). We found that the memory-based ACT-R model best explains the neural signatures for third person pronoun processing, primarily localized at the left middle temporal gyrus (LMTG) at around 320-350 ms after the onset of the pronouns. Our results suggest a domain-general mechanism for pronoun resolution that resembles memory retrieval.

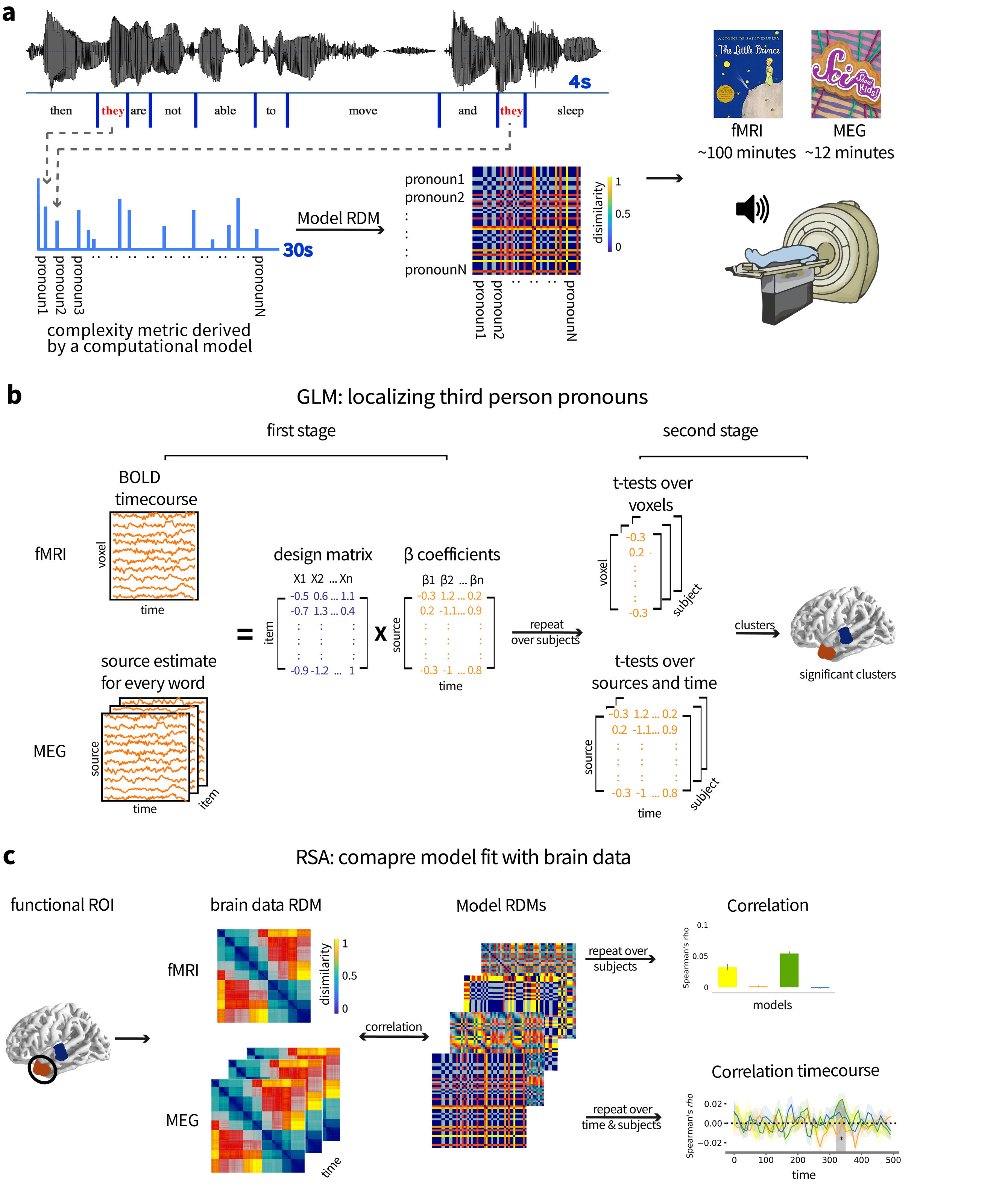


Figure 1: Schematic illustration of the analyses pipeline. **a** Complexity metrics for all the third person pronouns in the narratives are calculated based on different computational models for pronoun resolution. Participants listened to the narrative in the fMRI/MEG scanner. **b** GLM analyses to localize third person pronoun processing. **c** RSA analyses to compare model relatedness to brain activity pattern within the fROIs derived from the GLM analyses.

Results

Model comparisons. The Hobbs model[1] traverses the parsed syntactic tree of the sentence in a left-to-right, breadth-first order and searches for an antecedent that is matched in gender and number. It incorporates the locality constraints of the Binding Theory as it always searches for the antecedent in the left of the noun phrase (NP) and does not go below any NP or S(entence) Node on the tree. The Centering model[9] formalizes the Centering theory[4] for pronoun resolution. In the Centering framework, entities that link an utterance to others utterances are referred to as "centers". Centers of an utterance are ranked according to their relative prominence, which is mainly determined by the centers' grammatical roles. Pronouns are used when the most prominent centers of adjacent sentences are the same and form the preferred transition relation. Based on this assumption, the Centering algorithm tracks the relation between the centers in adjacent pairs of sentences and finds the antecedent-pronoun pair that has the most preferred transition relation. The memory-based ACT-R model[5] is specifically intended as a rigorous cognitive model for pronoun resolution. In the modular system of ACT-R[10], declarative memories of past events are stored as "chunks" in the buffer of the declarative module. The chunks have activation levels that determine the speed and success of their retrieval. The activation level is dependent on the frequency and recency effects of memory retrieval: the more often and more recent a chunk occurs, the more likely it is to be retrieved. In addition, spreading activation from other chunks can temporarily boost a chunk's activation: Chunks that are currently being processed spread activation to other, connected chunks in declarative memory. The ACT-R model for pronoun resolution uses the same primitives of the memory module in ACT-R. It calculates the activation levels of previous entities in the discourse context, and selects the most salient entity as the antecedent of the pronoun. The NeuralCoref model[6,7] learns the statistical pattern for clustering the mentions. At the core of the model are vector representations of each mention span, which are determined by the context surrounding each mention span and the internal structure of the span. The model uses bidirectional LSTMs to capture the contextual information with an attention mechanism to learn a notion of "syntactic headedness". Vector representation of each word is also crucial for the mention span, and contributed significantly to the full model performance as shown in the ablation experiment[6]. (see Figure 2 for an illustration of the four models applied to an example English sentence from "The Little Prince").

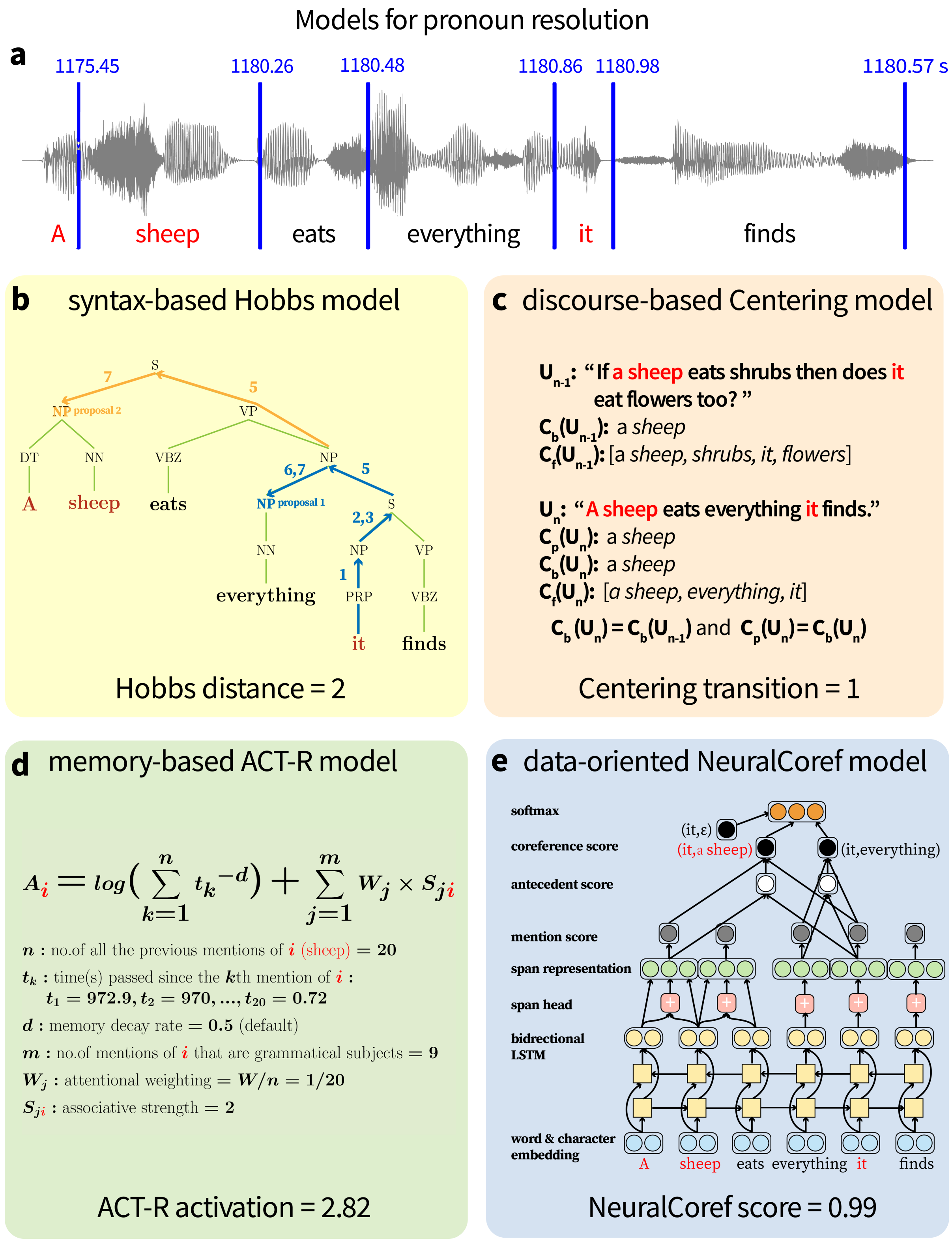

Models for pronoun resolution
a
1175.45
1180.26
1180.48
1180.86
1180.98
1180.57 s
A
sheep
eats
everything
it
finds
b syntax-based Hobbs model
S
7
5
NP proposal 2
VP
DT
NN
VBZ
NP
6,7
5
A
sheep
eats
NP proposal 1
S
2,3
NN
NP
VP
1
everything
PRP
VBZ
it
finds
Hobbs distance = 2
c discourse-based Centering model
U_n-1: "If a sheep eats shrubs then does it eat flowers too?"
C_b(U_n-1): a sheep
C_f(U_n-1): [a sheep, shrubs, it, flowers]
U_n: "A sheep eats everything it finds."
C_p(U_n): a sheep
C_b(U_n): a sheep
C_f(U_n): [a sheep, everything, it]
C_b(U_n) = C_b(U_n-1) and C_p(U_n) = C_b(U_n)
Centering transition = 1
d memory-based ACT-R model
A_i = log(Σ_{k=1}^{n} t_k^{-d}) + Σ_{j=1}^{m} W_j × S_ji
n : no.of all the previous mentions of i (sheep) = 20
t_k : time(s) passed since the kth mention of i :
t_1 = 972.9, t_2 = 970, ..., t_20 = 0.72
d : memory decay rate = 0.5 (default)
m : no.of mentions of i that are grammatical subjects = 9
W_j : attentional weighting = W/n = 1/20
S_ji : associative strength = 2
ACT-R activation = 2.82
e data-oriented NeuralCoref model
softmax
coreference score
antecedent score
mention score
span representation
span head
bidirectional LSTM
word & character embedding
(it,ε)
(it,a sheep)
(it,everything)
A
sheep
eats
everything
it
finds
NeuralCoref score = 0.99

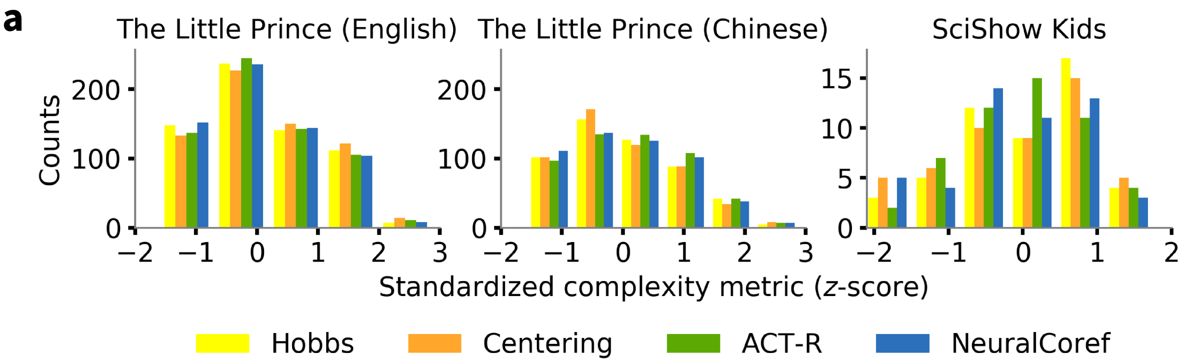
Figure 2: Illustration of the Hobbs[1], Centering[3], ACT-R[5] and NeuralCoref[6,7] algorithms using an example sentence in the English *The Little Prince*. **a** Waveform of the example sentence from the English audiobook. The blue numbers indicate the offset time in seconds for each word in the whole audiobook. **b** Hobbs model applied to the English example sentence. Blue arrows indicate the steps performed to get to the first proposed antecedent, and the orange arrows indicate the steps performed to get the second proposed antecedent. Hobbs distance is the number of proposals till the correct antecedent. **c** Centering model applied to the English example sentence. C_b equals C_{b-1} and C_p equals C_b, so the transition type is Continuing and the transition ordering is 1. **d** The formula for calculating the activation level for the antecedent "a sheep". **e** The architecture of the NeuralCoref model[6]. The NeuralCoref score is the softmax of the final layer.

Since the features used by each model vary across pronouns in the narratives, the models predict different "processing difficulty" for each pronoun. For example, pronouns that are far apart from its antecedents on a syntactic tree and are intervened by another candidate antecedent are hard for the Hobbs model; pronouns do not refer to the most prominent entity in the discourse context are hard for the Centering model; pronouns refer to entities that are mentioned only a few times a long time ago are hard for the ACT-R model, and pronouns surrounded by nouns with similar word meanings are hard for the NeuralCoref model (See Methods for a detailed description of the four models).

To connect properties of the four models to the observed brain data, we defined a "complexity metric" for each model to quantify how difficult it is for the model to find the correct antecedent. For the Hobbs model, we used the "Hobbs distance"[11], namely, the number of proposals that the Hobbs algorithm has to skip before the correct antecedent is found. For the Centering model, we used the rank of the transition type from the previous sentence to the current sentence containing the pronoun. For the ACT-R model, we used the negative of the activation level for the antecedent of each pronoun, and for the NeuralCoref model, we used the negative of coreferential probability of the antecedent for each pronoun. These complexity metrics allow us to estimate model-derived brain states for comparison against observed brain data.

Figure 3a shows the distribution of the standardized complexity metrics derived by the four models applied to all the third person pronouns in "The Little Prince" in English and Chinese and the "SciShow Kids" audio excerpts. We focused on third person pronouns as the three symbolic models are mainly concerned with third person pronoun resolution. In addition, the underlying neural mechanisms may differ for first, second and third person pronoun processing as first and second person pronouns have been suggested to mark proximity in space while third person pronouns are further away[12]. The distribution of the complexity metrics are very similar for the English and Chinese fMRI stimuli, where the complexity metrics are all right-skewed with more positive values of 2 or 3 z-scores. This suggest that texts contain more difficult pronouns for all the four models. For the MEG stimuli, the complexity metrics are left-skewed with more negative values, suggesting more easier pronouns for the models. This is expected as the "SciShow Kids" program is designed for kids and the narratives are kept simple and easy to understand.

To ensure that the four models can indeed predict the correct antecedents of the pronouns, we calculated the accuracy of the four models applied to all the third person pronouns in the narratives (see "Model performance" in Methods for details). We allowed some degree of ambiguity in the reference and permitted the correct answer to rank within a model's top 3 choices. This is the SUCCESS@N metric[13] where the gold answer occurs within a system's first N choices. All the models performed well with an accuracy at SUCCESS@3, with an accuracy well above or near 70% (Figure 3b).



b

Model	Accuracy (%)								
	The Little Prince (English)			The Little Prince (Chinese)			SciShow Kids		
	S@1	S@2	S@3	S@1	S@2	S@3	S@1	S@2	S@3
Hobbs	59.5	78.5	91.2	30.9	57.3	71.9	60.3	80.4	88.2
Centering	51.9	74.8	85.9	62.6	80.3	88.2	41.2	60.8	68.6
ACT-R	75.1	86.7	91.8	53.4	69.6	79.6	39.2	64.7	76.5
NeuralCoref	74.7	80.5	83.6	71.6	77.3	80.3	58.5	72.5	76.5

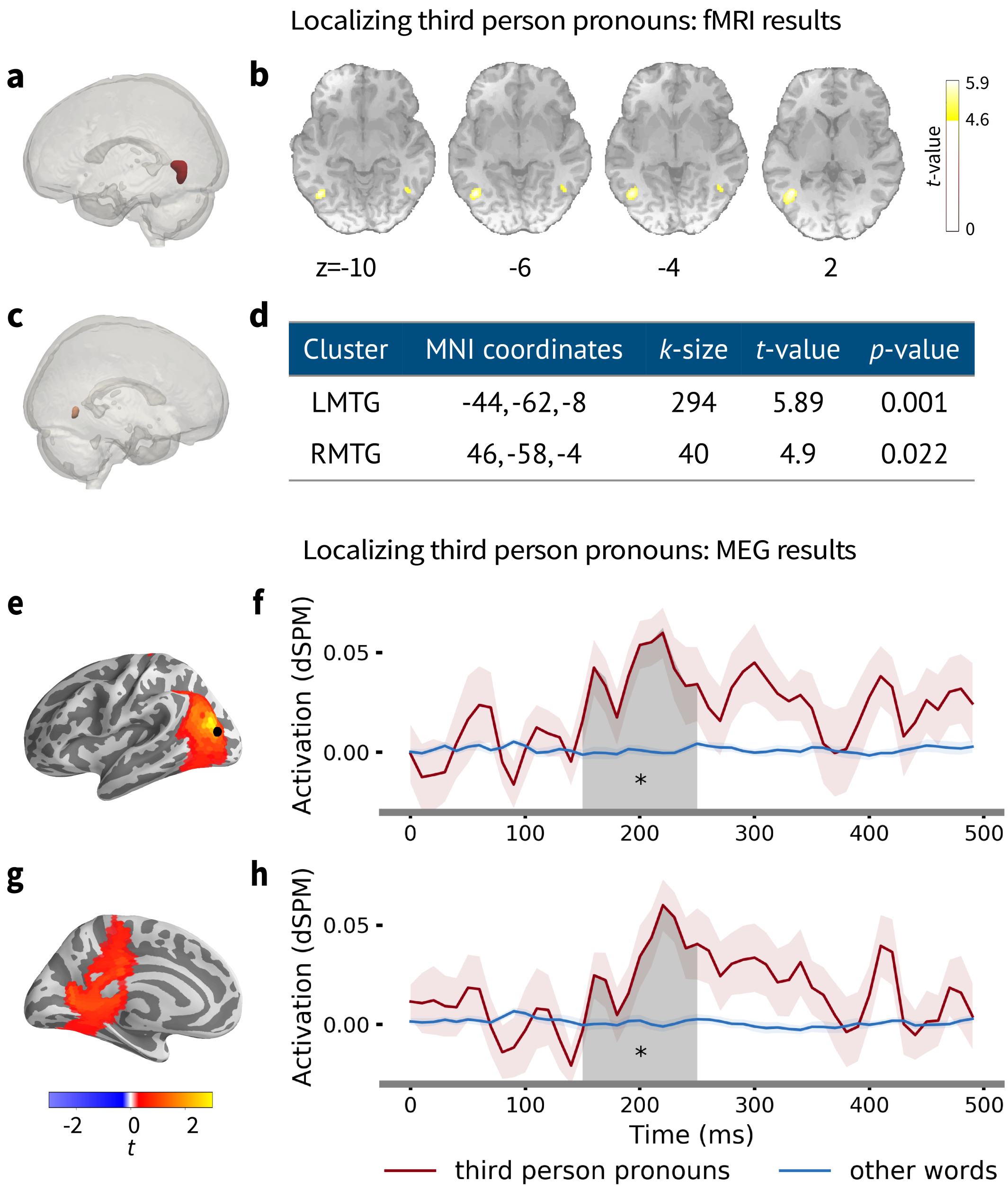
Figure 3: Comparison of the four models applied to the fMRI and MEG stimuli. **a** Distribution of the standardized model metrics. We took the negative of the ACT-R and the NeuralCoref metrics to indicate the processing difficulty of the pronouns, aligning with the Hobbs and the Centering metrics. All the metrics are z-scored. **b** The accuracy of the Hobbs, ACT-R and NeuralCoref model based on SUCCESS@N (N=1,2,3)[13], i.e., the proportion of the correct antecedent occurs within the model's first three choices.

Localizing third person pronoun processing. In order to compare the four models' relatedness to brain data, we need to first localize the regions involved in pronoun processing. Prior fMRI studies have suggested a number of regions relevant for pronoun processing, including the superior and middle temporal gyrus (STG, MTG), the inferior frontal gyrus (IFG), the angular gyrus (AG) and the Precuneous cortex (PC). The left STG, MTG and IFG have been shown to elicit increased activation with increased linear distance between pronouns and their antecedents[14,15], and the left AG and PC have been implicated in backward anaphora processing[16]. The MEG literature has mainly localized reference resolution to the medial parietal lobe[17,18]. However, no consensus has been reached on the exact location of third person pronouns, which are the focus of the current study.

To find brain regions showing increased activation during third person pronoun processing for both the English and Chinese participants in our fMRI study, we conducted a GLM analysis with a binary third person pronoun regressor, time-locked at the offset of each third person pronoun in the audiobook. We also included the root mean square (RMS) intensity and f0 of the audio, word rate and word frequency for each word in the story as control variables (Figure 1b; see "Localise brain regions for third person pronoun processing: fMRI data" in Methods for details). The results showed a significant cluster correlated with the occurrence of third person pronouns in the LMTG ($p = 0.001$ FWE, $k = 294$) and the RMTG ($p = 0.022$ FWE, $k = 40$; see Figure 4a-d).

The same regression model with the binary third person pronoun regressor was applied to the source-localized MEG data for each word at each source and each timepoint for each subject. Significant clusters for the β coefficient over space and time were identified by a cluster-based spatiotemporal permutation test[19] with 10,000 permutations (Figure 1b; see “Localise brain regions for third person pronoun processing: MEG data" in Methods for details). We found one significant cluster of 401 sources for the third person pronoun regressor from 150 to 250 ms (p=0.033) after the onset of the pronoun. The cluster covered regions including the left posterior temporal lobe (LPTL) and the left medial parietal lobe (Figure 4e,g). Figure 4f,h show the timecourses of the mean activation for third person pronouns and all other words averaged over the lateral and the medial parts of the cluster. The direction of the interaction is positive, which means that third person pronouns elicited higher activity compared to other words.

Both our fMRI and MEG results showed significant LMTG activity, consistent with previous literature[14,15,20] on pronoun resolution. Our MEG results showed additional activity in the left medial parietal lobe, which also replicated previous MEG results[17,18]. We extracted the LMTG and the RMTG clusters from the fMRI results as the functional regions of interests (fROIs) for further representational similarity analyses on the fMRI data. Similarly, we split the significant cluster from the MEG results into a lateral part and a medial part and used them as the fROIs for RSA on the MEG data.



Cluster	MNI coordinates	*k*-size	*t*-value	*p*-value
LMTG	-44,-62,-8	294	5.89	0.001
RMTG	46,-58,-4	40	4.9	0.022

Figure 4: GLM results for third person pronoun processing for the fMRI and MEG data. **a** The LMTG cluster derived from the fMRI data. **b** Coronal slices of the significant clusters. **c** The RMTG cluster derived from the fMRI data. **d** MNI coordinates, cluster size and their peak level statistics, thresholded at p<.05 FWE and k>20.**e** Lateral part of the significant cluster derived from the MEG data. **f** Timecourses of response for third person pronouns and all other words averaged over the lateral cluster. **g** Medial part of the significant cluster derived from the MEG data. **h** Timecourses of response for third person pronouns and all other words averaged over the medial cluster. Shaded region indicates the significant time window from 150-250 ms after the word onset ($p = 0.033$).

Comparing model predictions within the fROIs of the fMRI data. We performed an RSA[8] for each model within each fROIs to compare the model predictions with the brain activity patterns during third person pronoun resolution. RSA characterizes a representation in a brain or computational model by the representational dissimilarity matrix (RDM) of the brain activity patterns or the model predictions. A model is tested by comparing the RDM it predicts to that of a measured brain region. We computed the Hobbs distance, the Centering transition, the

ACT-R activation level, and the NeuralCoref score for each pronoun in the English and Chinese fMRI stimuli. We then constructed the representational dissimilarity matrix (RDM) for each model using the euclidean distance between each pronoun‘s model predictions. Each cell of the RDM represents the dissimilarity of a pronoun compared to other pronouns based on model predictions (Figure 1a). Figure 5a,c show the four model RDMs for the English and Chinese stimuli, ranked transformed and scaled into [0,1]. Spearman‘s rank correlation (*rho*) revealed low correlation among the model RDMs for both populations, with the highest *rho* value being 0.18 for the Centering and NeuralCoref RDMs of the English stimuli (Figure 5b,d). We then correlated each model RDM with the fMRI data RDM within each fROI for each subject within each group. The fMRI data RDMs are computed as 1 minus Pearson‘s *r* correlation among all the fMRI scans aligned with the occurrence of a third person pronouns in the audiobook. Each cell in the fMRI data RDM reflects dissimilarity of brain activity patterns among the fMRI scans containing a pronoun. Statistical significance was tested using a one-sided signed-rank test[21] across each subject‘s correlation values. (see “RSA within the fROIs: fMRI data” in Methods for details).

The correlational results showed similar patterns for both the English and Chinese speakers (Figure 5e,f). For English speakers, the ACT-R model showed the highest Spearman‘s rank correlation with both the LMTG and RMTG activity patterns, averaged across subjects ($rho = 0.054, p < .0001$ and $rho = 0.053, p < .0001$, respectively). The Hobbs model was also significantly related to both fROI activities ($rho = 0.032, p < .0001$ and $rho = 0.031, p < .0001$, respectively). The Centering and the NeuralCoref model RDMs were not significantly correlated with either of the fROI activities. For Chinese speakers, the ACT-R model also had the highest mean correlation with both the brain regions ($rho = 0.084, p < .0001$ and $rho = 0.093, p < .0001$, respectively). The Hobbs model was also significant for both fROIs with a *rho* value of 0.023 ($p < .0001$) and 0.031 ($p < .0001$), respectively. The NeuralCoref model showed a significant but very low correlation with the LMTG activity ($rho = 0.0008, p = 0.02$). The Centering model was not significant for either of the fROI activities. A two-sample *t*-test between the English and Chinese speakers revealed a significantly higher correlation of the ACT-R models to the brain data in Chinese ($t(82) = 4.81, p < .0001$), suggesting a better fit of the ACT-R model for pronoun resolution in Chinese.

Given that the four models focus on different aspects of pronoun resolution, it may well be the case that they are distinctively correlated with different brain regions in a broad network for pronoun processing. To search for regions associated with each model, we conducted a searchlight RSA within a mask covering regions that have been previously reported for pronoun processing, including the STG, MTG, IFG, AG and PC[14,15,16,17,22] (see “Searchlight RSA: fMRI data” in Methods for details).

The searchlight results showed a significant correlation between the ACT-R model and all the brain regions in the searchlight mask for both English and Chinese speakers. For English speakers, the peak correlation was localized at the left and right IFG (MNI[-52,28,12], $p < .0001$ FWE and MNI[42,6,42], $p < .0001$ FWE, respectively). Chinese speakers had a peak correlation for the ACT-R model at the LIFG (MNI[-44,12,28], $p < .0001$ FWE). Group comparison between the correlation maps revealed a significant cluster at the LIFG where the ACT-R model has a better fit for the Chinese data than for the English data (MNI[-44,12,28], $p < .0001$ FWE, $k = 51$). The other three models did not show any significant correlation with the BOLD response patterns in the searchlight mask for both English and Chinese speakers.

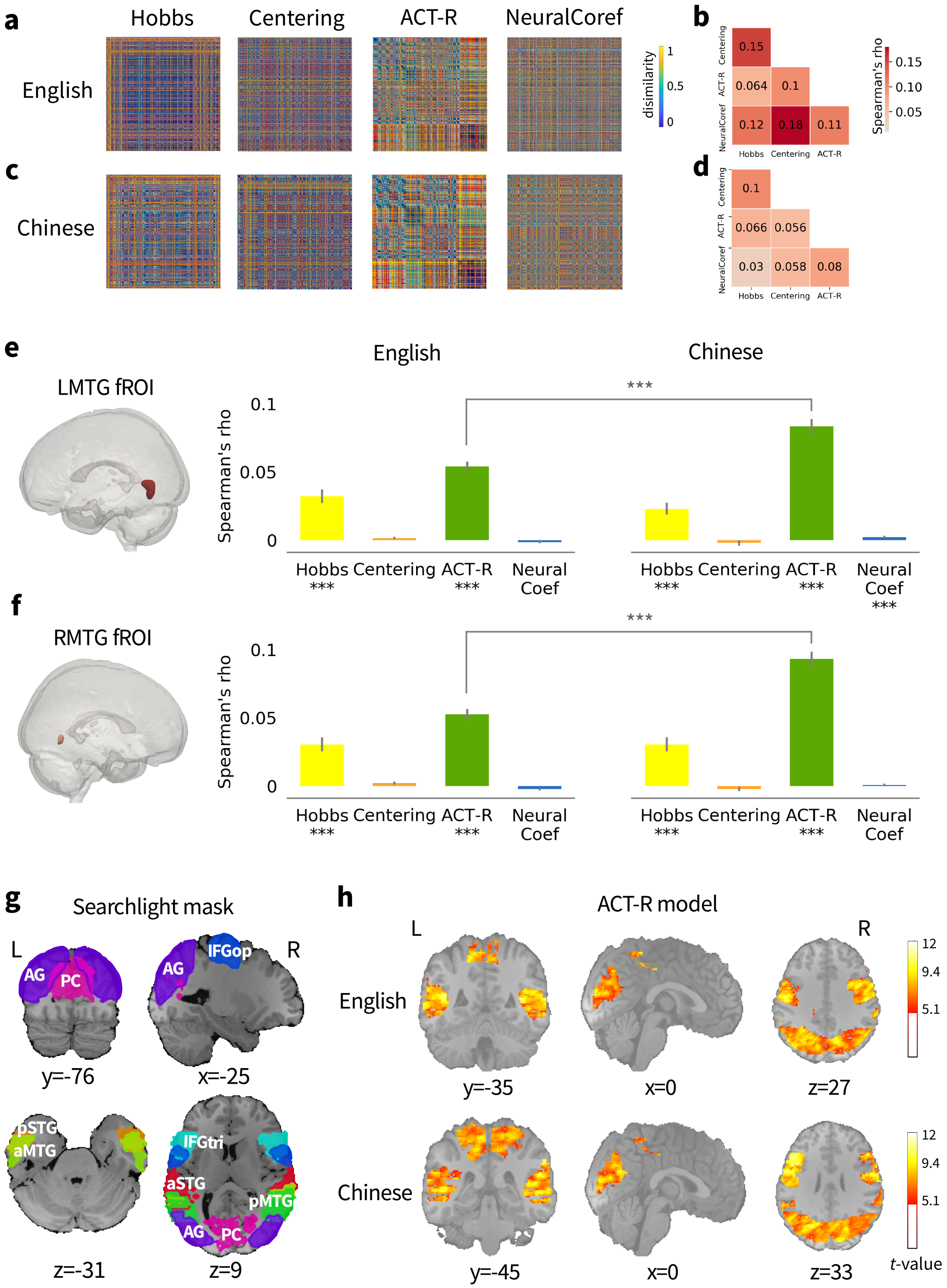


Figure 5: ROI-based and searchlight RSA results for the fMRI data. **a** The RDMs for the Hobbs, Centering, ACT-R and NeuralCoref model in English. Each RDM was separately rank-transformed and scaled into [0,1]. **b** Spearman's rank correlation matrix for the four model RDMs in English. **c** The model RDMs in Chinese. **d** Spearman's rank correlation matrix for the four model RDMs in Chinese. **e** Relatedness of the four model RDMs to the brain data RDM within the LMTG fROI, averaged across subjects in each group. **f** Relatedness of the four model RDMs to the fMRI data RDM within the RMTG fROI, averaged across subjects in each group. * below the model names indicates significant correlation. * above the bars indicates significant group difference. FDR correction was applied for multiple comparisons across fROIs and models. *** p<.0001.**g** The selected anatomical regions based on the Harvard-Oxford cortical atlas[23] for searchlight RSA. **h** Regions showing significant correlation for the ACT-R model from the searchlight analyses. Significance was determined by a one-sided t-test at the group level with FWE $p < .005$ and cluster size > 20. aMTG/pMTG: anterior/posterior Middle Temporal Gyrus; IFGop/IFGtri: Inferior Frontal Gyrus pars opercularis/pars triangularis; aSTG/pSTG: anterior/posterior Superior Temporal Gyrus; AG: Angular Gyrus; PC: Precuneous Cortex.

Whole-brain GLM analyses for the fMRI data. Our multivariate RSA methods selected the fMRI scans that are aligned with the occurrence of third person pronouns in the audiobook, yet given the fMRI time resolution of 2 seconds per volume, the words surrounding the pronouns are likely included in the selected scans too. Although we tried to remove the effects of other words by regressing out the intensity, f0, word rate and word frequency effects, there is still a concern of contamination from other words. To complement the multivariate approach, we also conducted a univariate GLM analysis with the four models' predictions as regressors. The GLM approach has been well-established to examine neuro-computational models of language processing in prior fMRI studies using naturalistic stimuli[24,25,26]. Same with the GLM analysis for localizing third person pronouns, we modeled the timecourse of each voxel's BOLD signals for each of the nine sections by the model regressors, time-locked at the offset of each third person pronoun in the audiobook (see "Whole-brain GLM analyses: fMRI data" in Methods for details).

The results of the GLM analyses were shown in Figure 6. Consistent with the searchlight RSA results, the ACT-R model was significant for a temporal-frontal network for both English and Chinese speakers, while the other three models were associated with much smaller and isolated clusters across the two groups. For English speakers, the peak clusters for the ACT-R activation were at the LIFG ($p < .0001$ FWE, $k = 4440$), LMTG ($p < .0001$ FWE, $k = 2451$) and RMTG ($p < .0001$ FWE, $k = 547$). For Chinese speakers, the peak clusters were at the LAG ($p < .0001$ FWE, $k = 517$), LMTG($p < .0001$ FWE, $k = 281$), LSFG ($p < .0001$ FWE, $k = 426$), and LIFG ($p = 0.001$ FWE, $k = 72$; see Figure 6c). Direct comparison of the contrast maps for the ACT-R model showed no significant cluster between the English and Chinese groups. The Hobbs model was correlated with a significant cluster in the RMTG ($p < .0001$ FWE, $k = 428$) for English speakers and the LSMA ($p = 0.001$ FWE, $k = 109$) for Chinese speakers (Figure 6a). English speakers showed significant higher activation than Chinese speakers in the RMTG ($p < 0.0001$ FWE, $k = 201$). The Centering model was associated with two small clusters in the RSFG ($p < .0001$ FWE, $k = 74$) and the PC ($p = 0.005$ FWE, $k = 60$) for English speakers. No significant cluster was found for the Centering model for Chinese speakers (Figure 6b). Group comparison revealed greater activation for the Centering model in the PC ($p = 0.01$ FWE, $k = 68$) for English speakers. The NeuralCoref model was correlated with significant clusters in the LSTG ($p < .0001$ FWE, $k = 1273$), RMTG ($p < .0001$ FWE, $k = 305$), LAG ($p = 0.002$ FWE, $k = 120$) and LSFG ($p = 0.002$ FWE, $k = 134$) for English speakers. Chinese speakers also showed a significant cluster in the LSTG for the NeuralCoref model ($p < .0001$ FWE, $k = 144$; see Figure 6d). Supplemental Table 1 lists the MNI coordinates and the statistics for the peak clusters for the four models.

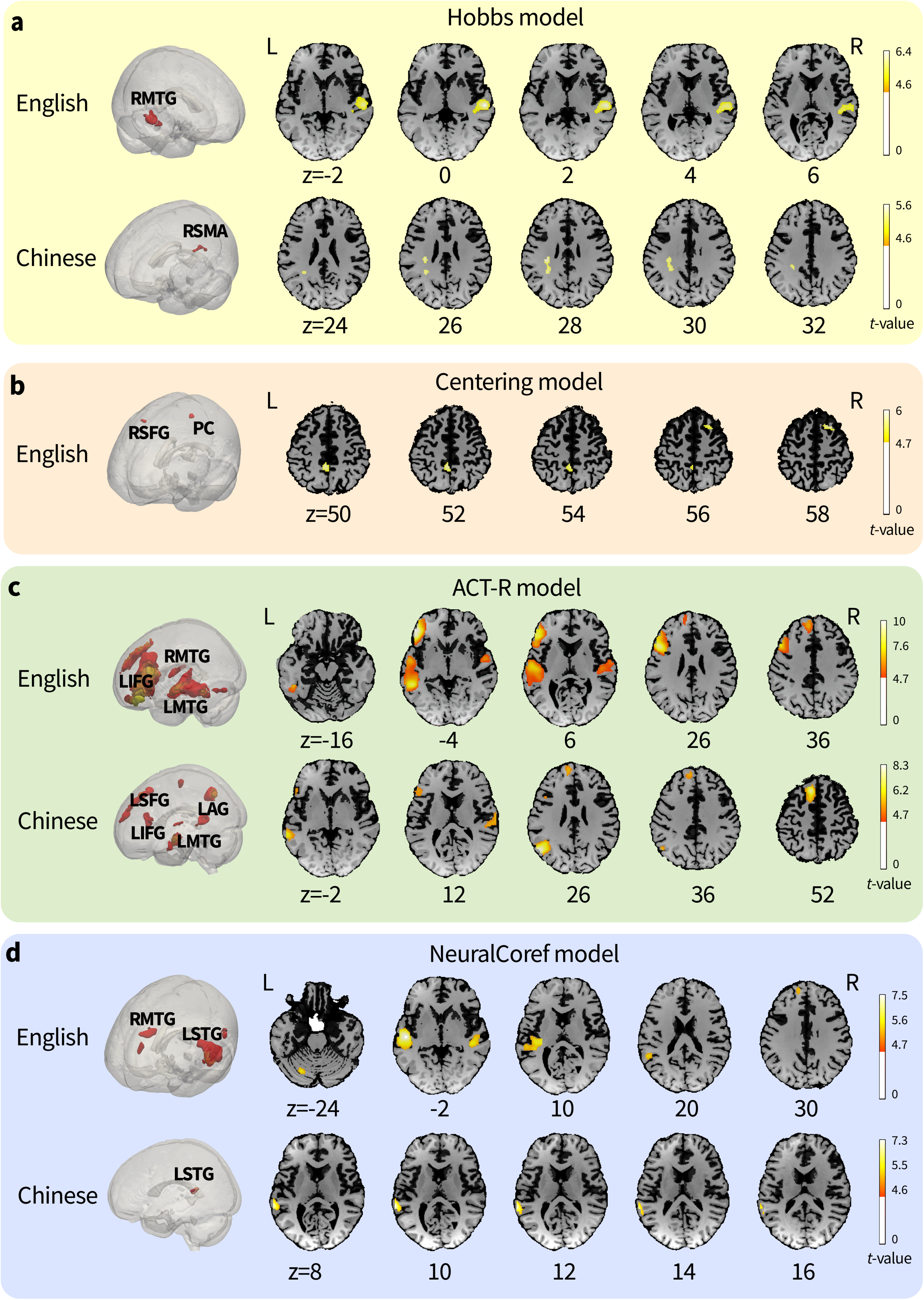


Figure 6: GLM results with the four models as regressors for the fMRI data. **a** 3D visualisation and the coronal slices showing the significant clusters for the Hobbs model in English and Chinese. **b** 3D visualisation and the coronal slices showing the significant clusters for the Centering model in English. **c** 3D visualisation and the coronal slices showing the significant clusters for the ACT-R model in English and Chinese. **d** 3D visualisation and the coronal slices showing the significant clusters for the NeuralCoref model in English and Chinese.

Comparing model predictions within the fROIs of the MEG data. We computed the Hobbs distance, the Centering transition, the ACT-R activation level and the NeuralCoref score for each pronoun and calculated the model RDMs using the euclidean distance metric (see Figure 7a,b for the model RDMs and the correlation matrix between the model RDMs). We then extracted the source estimates for each pronoun within the lateral and medial fROIs, and calculated the MEG data RDMs as 1-Pearson'r correlation between each source estimate. We calculated the Spearman's rank correlation between each model RDM with the MEG data RDM within each fROI at each timepoint. Statistical significance was tested using a cluster-based permutation test over time[19] (Figure 1c; see "RSA within the fROIs: MEG data" in Methods for details).

We observed a significant temporal cluster for the ACT-R model from 320 to 350 ms after the onset of the pronouns for both the lateral fROI ($p = 0.049$) and the medial fROI ($p = 0.038$). The other three models were not significantly correlated with the MEG data patterns within the two fROIs (see Figure 7).

To further examine the models' relatedness to the MEG data in other brain regions, we also conducted a searchlight RSA within the same mask used for the fMRI searchlight RSA (see "Searchlight RSA: MEG data" in Methods for details). Figure 8 showed the t-values of the statistical tests on the correlation maps for each model on the left hemisphere, thresholded at t>1. Similar to the ROI-based RSA results, the ACT-R model showed similar posterior LMTG and LPC activation at around 300-400 ms. The Hobbs model showed the highest correlation at the left prefrontal cortex at around 200 ms, the Centering model showed the highest correlation with the LPC at around 200 ms, and the NeuralCoref model showed the highest correlation at the meidal prefrontal cortex at around 100 ms. Although none of these clusters survived the cluster-based permutation test with the cluster-level threshold of $p < .05$, the LMTG cluster for the ACT-R model has the smallest p-value of 0.29.

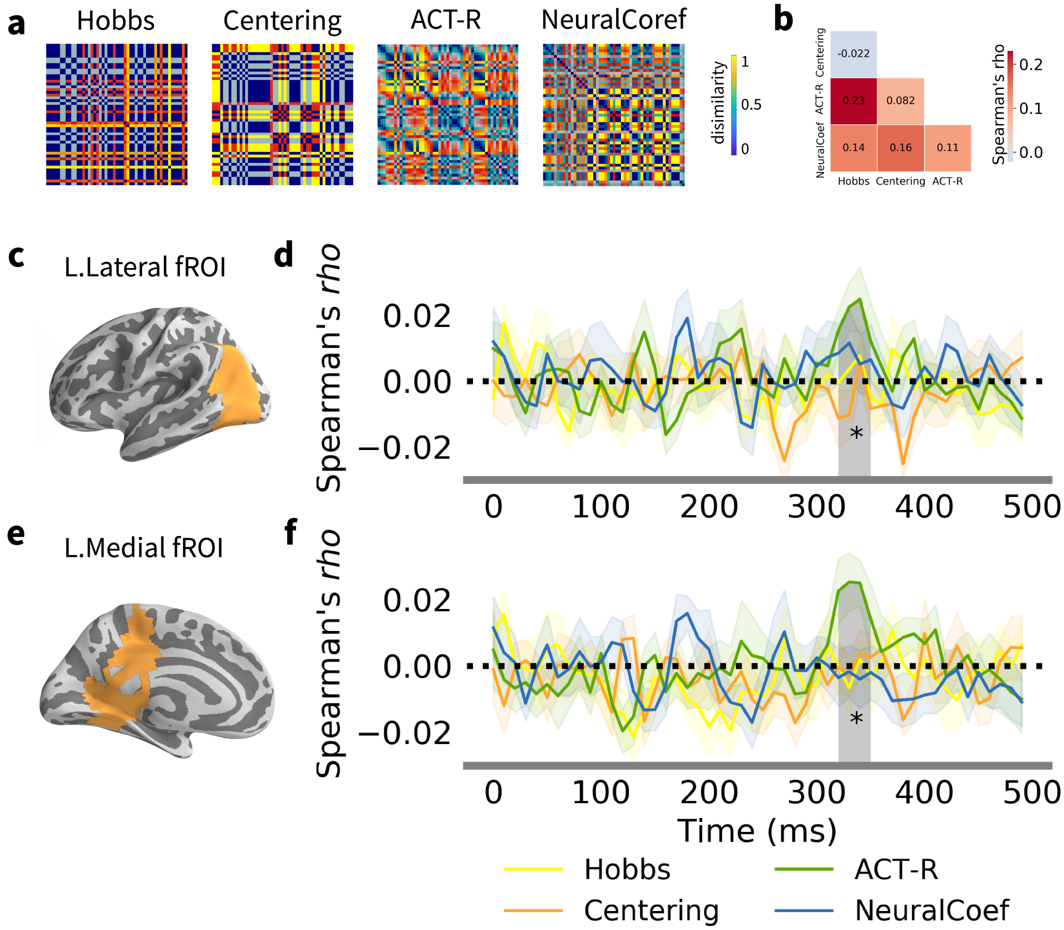


Figure 7: ROI-based RSA results for the MEG data. **a** The RDMs for the Hobbs, Centering, ACT-R and NeuralCoref model metrics for the MEG stimuli. Each RDM was separately rank-transformed and scaled into [0,1]. **b** Spearman's rank correlation matrix for the four model RDMs. **c** Functional ROI in the left lateral lobe derived from the regression analyses for third person pronouns. **d** Timecourse of the relatedness of the four model RDMs to the brain data RDM within the lateral temporal fROI, averaged across subjects in each group. ACT-R model is significantly correlated with the MEG data pattern in the lateral fROI from 320-350 ms ($p = 0.049$). **e** Functional ROI in the left medial wall derived from the regression analyses for third person pronouns. **f** Timecourse of the relatedness of the four model RDMs to the brain data RDM within the lateral temporal fROI, averaged across subjects in each group. ACT-R model is significantly correlated with the MEG data pattern in the medial fROI from 320-350 ms ($p = 0.038$). Shaded region indicates significant temporal cluster. * p<.05.

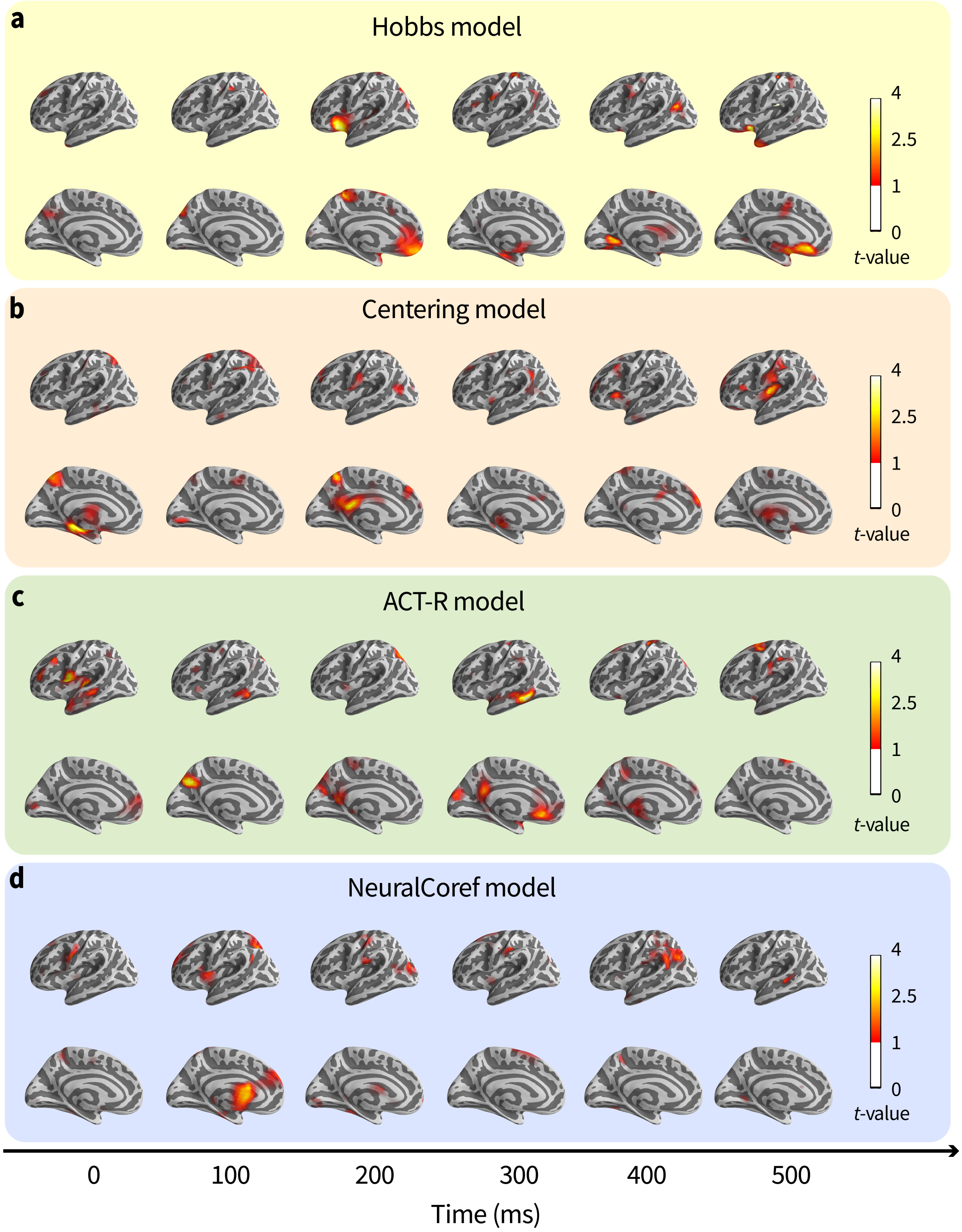


Figure 8: Searchlight RSA results for model RDMs and the MEG data RDMs. **a** Spatiotemproral pattern of the *t*-values from the permutation *t*-test for correlation maps of the Hobbs model RDM and the MEG data across subjects in the left hemisphere, thresholed at $t > 1$. **b** Spatiotemproral pattern of the *t*-values for the Centering model. **c** Spatiotemproral pattern of the *t*-values for the ACT-R model. **d** Spatiotemproral pattern of the *t*-values for the NeuralCoref model.

Discussion

Computational models for pronoun resolution provide a viable way of specifying complex and detailed theories of the underlying cognitive process. Consequently, they make quantitative predictions that can be rigorously tested against human brain activity. Here, we tested three knowledge-based symbolic models and one data-oriented neural network model for pronoun resolution against both fMRI and MEG data. Our results all favor the ACT-R model[5]: For the fMRI data, the ACT-R model showed the highest correlation with the BOLD response patterns within the bilateral MTG fROIs, and the searchlight RSA revealed significant correlation of the model against all the brain regions previously reported for pronoun processing; the GLM analyses also showed a network of regions associated with the ACT-R model; for the MEG data, the ACT-R model is the only model that showed significant correlation with the source-localized MEG data within the LPTL and the LPC fROIs. The significant time window is around 320-350 ms after the onset of the pronouns.

Built using the primitives of the memory module in the cognitive architecture ACT-R[10], the ACT-R model views pronoun resolution as the process of retrieving the most salient entities from declarative memory. It incorporates the frequency and recency effects of memory decay, with a boost of activation spread from previous mentions of the entities that occupy the subject position of a sentence. Although the Hobbs and the NeuralCoref model also incorporate a notion of recency, and the Centering model contains the feature of subjecthood, the parameters for the ACT-R formula was directly developed using relevant fMRI data[10]. Our correlational results suggest that this algorithm indeed traces the states of computation of the brain better than other non-brain-inspired models.

Both the fMRI and the MEG data showed significant correlation of the ACT-R model with the LMTG, which has been previously associated with biological and syntactic gender processing during pronoun resolution[14,15,20]. Lesions in LMTG also led to aphasic patients with selective difficulty to access nouns[27,28]. We also showed an LIFG activation for the ACT-R model under the whole-brain GLM analyses on the fMRI data, consistent with previous findings that longer distance between a back anaphora and its referent leads to increased activity in the LIFG. In addition, the original proposal from ACT-R also states that the memory module is associated with the left prefrontal region[10].

The NeuralCoref model were associated with significant LSTG activity for both English and Chinese under the GLM analyses for the fMRI data, but the RSA results did not show significant correlation of the model with either the fMRI or MEG activity patterns. Deep learning models have led to significant advances in many aspects of natural language processing, yet many of these models are not intended to match human cognitive process. Here we show that memory retrieval principles formalized in the ACT-R model may be incorporated into the architecture of of a neural coreference model that is based on human cognition.

To sum up, we show that computational models can be leveraged to understand the fine details of the neural mechanisms underlying pronoun resolution. By testing computational models rigorously against neuroimaging data, we also hope to provide insights on designing machines that learn and think like human. We call for a joint force of cognitive neuroscience and artificial intelligence to explicate the intricate details of the human mind.

Methods

Participants. The English fMRI data were taken from a published study[26], collected at the same time with the Chinese data. Participants were 49 young adults (30 female, mean age =

21.3, SD=3.6) with no history of psychiatric, neurological or other medical illness that might compromise cognitive functions. They self-identified as native English speakers, and strictly qualified as right-handed on the Edinburgh handedness inventory[29]. All participants were paid, and gave written informed consent prior to participation, in accordance with the IRB guidelines of Cornell University.

Chinese participants of the fMRI study are 35 young adults (15 female, mean age=19.3, SD=1.6) with with normal hearing and no history of psychiatric, neurological or other medical illness that might compromise cognitive functions. They self-identified as native Chinese speakers, and strictly qualified as right-handed on the Edinburgh handedness inventory[29]. All participants were paid, and gave written informed consent prior to participation, in accordance with the IRB guidelines of Jiangsu Normal University.

Participants for the MEG study were 13 young adults (7 female, mean age=19.9, SD=1.3) with normal hearing and no history of psychiatric, neurological or other medical illness that might compromise cognitive functions. They self-identified as native English speakers, and strictly qualified as right-handed on the Edinburgh handedness inventory. All participants were paid, and gave their written informed consent prior to participation, in accordance with New York University Abu Dhabi IRB guidelines.

Stimuli and annotation. The English fMRI stimulus is an audiobook version of Antoine de Saint-Exupéry's *The Little Prince*, translated by David Wilkinson and read by Nadine Eckert-Boulet. The Chinese fMRI stimulus is a Chinese translation of *The Little Prince*[30], read by a professional female Chinese broadcaster hired by the experimenter. The MEG stimulus is an audio excerpt taken from the YouTube channel "SciShow Kids"[31]. It consists of 4 short audios that introducing scientific fun facts to kids: "Use your brain!", "Why do we get dizzy?", "Why do we need sleep?" and "Why do we get goosebumps?"

All mentions in the three texts were first identified using the Stanford Named Entity Recognizer (NER)[32]. They were then manually checked and linked with their coreferential mentions using the annotation tool brat[33]. Supplemental Figure 1 demonstrates sample annotations for the three texts. We identified 4882 mentions in the English fMRI stimulus, 4732 mentions in the Chinese fMRI stimulus, and 701 mentions in the MEG stimulus. We further removed possessives, reflexives, cleft and extraposition "it", pleonastic "it"and pronouns with sentential antecedents. The final English fMRI stimulus contain 647 third person pronouns, the Chinese fMRI stimulus contains 524 third person pronouns, and the MEG stimulus contains 51 third person pronouns (Supplemental Figure 2).

Speech segmentation. Word boundaries in the fMRI audios were identified and aligned to the transcript using the Penn Phonetics Lab Forced Aligner[34] and were manually checked by native English and Chinese speakers. The MEG audio was aligned to the transcript using the Forced Alignment and Vowel Extraction (FAVE)[35] and were manually checked by native English speakers.

The Hobbs model. The Hobbs model for pronoun resolution[1] depends on a syntactic parser plus a morphological gender and number checker. The input to the Hobbs model includes the target pronoun and the parsed trees for the current and previous sentences. The model searches for a gender and number matching antecedent by traversing the trees in a left-to-right, breadth-first order, that is, it starts at the tree root and explores the neighboring nodes at the present depth prior to moving on to the nodes at the next depth level. If no candidate antecedent is found in the current tree, the algorithm searches on the preceding sentence in the same order. The steps of the Hobbs algorithm are as follows:

1 Begin at the NP node immediately dominating the pronoun.

2 Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.

3 Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node that is encountered which has an NP or S node between it and X.

4 If node X is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as antecedent. If X is not the highest S node in the sentence, continue to step 5.

5 From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.

6 If X is an NP node and if the path p to X did not pass through the $\bar{N}$ node that X immediately dominates, propose X as the antecedent.

7 Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.

8 If X is an S node, traverse all branches of node X to the right of path p in a left-to-right. breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.

9 Go to step 4.

The Hobbs algorithm conforms to the Binding Theory as it always searches the antecedent in the left of the NP (Principle B: Step 3) and do not go below any NP or S node encountered (Principle A: Step 8). When applied to the Chinese fMRI stimulus, the Hobbs no longer contains a gender and number agreement checker because pronouns in spoken Chinese do not distinguish gender and Chinese NPs usually do not mark plurals.

We use the “Hobbs distance"[11] metric to represent the processing complexity of the pronouns derived by the Hobbs model. Hobbs distance refers to the number of proposals that the Hobbs algorithm has to skip, starting backwards from the pronoun, before the correct antecedent NP is found. Figure 2c,d illustrate the Hobbs model for one example sentence in English and Chinese. For the English sentence, the model first proposes the noun phrase (NP) “everything" as the antecedent of “it", which is incorrect, so the Hobbs distance is 2.

The Centering model. The Centering model for pronoun resolution[3] (also known as the BFP algorithm) formalizes the Centering Theory[4], which argues that certain entities mentioned in an utterance were more central than others, and this property leads the speaker to use pronouns. In the Centering framework, entities that link an utterance to others utterances are referred to as “centers”. Centers of an utterance are ranked according to their relative prominence, which is mainly determined by the centers‘ grammatical roles. In particular, SUBJECTS of a sentence ranks higher than OBJECTS, and OBJECTS rank higher than other grammatical roles. Each utterance (U_n) has a set of forward-looking centers (C_f) and a single backward-looking center (C_b). $C_f(U_n)$ contains all the entities in U_n and C_b is the highest-ranked entity among the entities in the previous utterance ($C_f(U_{n-1})$). The transition relations between the forward- and backward-looking centers in an adjacent pair of sentences are classified into three types: CONTINUING, RETAINING and SHIFTING. In the CONTINUING transition, propositions of the current entity are maintained, that is to say, $C_b(U_n)$ is the same entity as the backward-looking center of the previous utterance ($_Cb(U_n) = C_b(U_{n-1})$), and $C_b(U_n)$ is also the preferred center of the current utterance ($C_p(U_n)$), i.e., the highest-ranked entity in $C_f(U_n)$ ($C_b(U_n) = C_p(U_n)$). In the RETAINING transition, a related entity is introduced to the context, thus $C_b(U_n)$ is the same as $C_b(U_{n-1})$, but it is not the highest-ranked entity in $C_f(U_n)$ ($C_b(U_n) = C_b(U_{n-1})$ and $C_b(U_n) \neq C_p$). In the SHIFTING transition, a new entity becomes the center of the discourse, therefore, $C_b(U_n)$ is not the same entity as $C_b(U_{n-1})$ ($C_b(U_n) \neq C_b(U_{n-1})$). For a discourse segment to be coherent, CONTINUING transitions are preferred over RETAINING transitions, which are preferred over SHIFTING transitions. Frequent SHIFTING leads to a lack of discourse coherence and substantially affects the processing demands made upon a hearer during discourse comprehension.

The Centering Theory claims that pronominalization serves to increase discourse coherence and eases the hearer's processing difficulty of inference. Based on this assumption, the Centering model[3] tracks the relation between the forward- and backward-looking centers in adjacent pairs of sentences and finds the antecedent-pronoun pair that has the highest-ranked transition types. The model further divides the SHIFTING transition into SHIFTING where $C_b(U_n) \neq C_b(U_{n-1})$ and $C_b(U_n) = C_p$, and SHIFTING-1, where $C_b(U_n) \neq C_b(U_{n-1})$ and $C_b(U_n) \neq C_p$. The coherence ordering of the transition types is CONTINUING > RETAINING > SHIFTING > SHIFTING. The algorithm consists of three basic steps. In the first step, it constructs all possible $C_b - C_f$ pairs for the pronoun in the current utterance (U_n). These pairs are called "anchors" and they represent all the coreferential relationships available for this utterance. Next, the model filters the anchors based on the Centering rule that C_b must be pronominalized if any C_f is pronominalized. Finally, the algorithm ranks the remaining pairs by the transition ordering and select the pair that has the most preferred transition types. Rank of the entities in C_f is determined by their grammatical roles, which is parsed using the Stanford dependency parser for English[36] and Chinese[37].

We used the rank of the transition types for each correct antecedent-pronoun pair generated by the Centering model to indicate the processing difficulty of the pronouns. Figure 2c illustrates how the model classifies the transition type for the example discourse segment in "The Little Prince". The current utterance (U_n) "A sheep eats everything it finds" has a set of $C_f(U_n)$: ("a sheep", "everything", "it"); the preferred center ($C_p(U_n)$) is "a sheep" as it is the subject of the sentence. The backward-looking center of the current ($C_b(U_n)$) and the previous utterance ($C_b(U_{n-1})$) is also "a sheep" as it is the subject of the U_{n-1} (highest-ranked entity in $C_f(U_{n-1})$). Since $C_b(U_n) = C_b(U_{n-1})$ and $C_b(U_n) = C_p(U_n)$, the transition type is CONTINUING and the rank is 1.

The ACT-R model. The ACT-R model for pronoun resolution[5] uses the same primitives of the memory module in the cognitive architecture ACT-R[38]. The formula for the activation level for the antecedent i of a pronoun is as follows:

$$A_i = log(\sum_{k=1}^{n} {t_k}^{-d}) + \sum_{j=1}^{m} W_j \times Sji$$

The first part of the equation $log(\sum_{k=1}^{n} t_k{}^{-d})$ computes the inherent strength, or the base-level activation of entity i, which reflects the past usage of entity i in the text. t_k is the time passed since the kth mention of i, and each mention decays over time as a negative power function $t_k{}^{-d}$. The parameter d is set to 0.5 as the default value in ACT-R based on a range of experiments to model human performance in memory retrieval tasks[38]. Different mentions of the entity i adds up to reflect the effect of practice. We calculated the base-level activation of each pronoun in English and Chinese based on the offset time of each pronoun and their previous mentions in the whole audio.

The second part of the equation $\sum_{j=1}^{m} W_j \times S_{ji}$ reflects the associative activation that entity i receives from the mentions of i which is a subject of its sentence. S_{ji} is the strength of association reflecting how much the presence of each subject mention j of entity i makes i more salient. The value of S_{ji} is set to 2 in our implementation. W_j is the attentional weighting which equals to W/n where n is the number of all the previous mentions of i, as the total value of associative activation cannot be infinite. The attentional weight W is set to 1. Subjecthood of each mention in the English and Chinese texts was annotated using the Stanford dependency parser[36,37].

The effects of frequency and recency are folded into the calculation of the base activation for antecedent i, such that the more mentions it has, and the more recent the mentions occur, the higher the base activation. Conversely, if antecedent i has been mentioned only once, or if its last mention was a long time ago, its activation level will be low, and it will rank lower on the activation list for all the candidate antecedents. Subjecthood of the previous mentions of antecedent i gains an associative activation in addition to the base activation. Overall, the amount of activation value of an entity in the discourse context is computed based on recency, frequency and grammatical role of the entity, and the entity that has the highest activation level is predicted to be the antecedent of the pronoun. Figure 2d shows how the ACT-R activation level for a pronoun in the English example sentence is calculated.

The NeuralCoref model. The NeuralCoref model[6,7] is an end-to-end coreference resolution system that predicts all clusters of coreferential mentions given a text document and its speaker and genre metadata. The model considers all possible spans in a document D containing T words as potential mentions. The total number of possible text spans in D is $N = T(T+1)/2$. The task for the model is to assign an antecedent y_i for each span i with a start and end index of $START(i)$ and $END(i)$ for $1 \leq i \leq N$. The possible assignments of each y_i is $Y(i) = \{\epsilon, 1, ..., i-1\}$. The model then learns a conditional probability distribution over the possible antecedents using the pairwise score $s(i, y_i)$ for each antecedent-span pair:

$$
\begin{aligned}
P(y_1, ..., y_N|D) &= \prod_{i=1}^{N} P(y_i|D) \\
&= \prod_{i=1}^{N} \frac{exp(s(i, y_i))}{\sum_{y' \in Y(i)} exp(s(i, y'))}
\end{aligned}
$$

The architecture of the model can be divided into two parts. In the first part, the model encodes every word in its context using bidirectional LSTMs[39]:

$$
\begin{aligned}
f_{t,\delta} &= \sigma(W_f[x_t, h_{t+\delta,\delta}] + b_i) \\
o_{t,\delta} &= \sigma(W_o[x_t, h_{t+\delta,\delta}] + b_o) \\
\tilde{c}_{t,\delta} &= tanh(W_c[x_t, h_{t+\delta,\delta}] + b_c) \\
c_{t,\delta} &= f_{t,\delta} \circ \tilde{c}_{t,\delta} + (1 - f_{t,\delta}) \circ c_{t+,\delta} \\
h_{t,\delta} &= o_{t,\delta} \circ tahn(c_{t,\delta}) \\
x_t^* &= [h_{t,1}, h_{t,-1}]
\end{aligned}
$$

The model then assigns weights to the word vectors to represent the notion of syntactic head using an attention mechanism[40]:

$$
\begin{aligned}
\alpha_t &= w_\alpha \cdot FFNN_\alpha(x_t^*) \\
a_{i,t} &= \frac{exp(\alpha_t)}{\sum\limits_{k=START(i)}^{END(i)} exp(\alpha_k)} \\
\hat{x}_i &= \sum_{t=START(i)}^{END(i)} a_{i,t} \cdot x_t
\end{aligned}
$$

The context-dependent word embeddings and their weighted sum are concatenated to produce the span representations. In the second part, the model assigns a mention score to each span representation and a antecedent score to each antecedent-span pair via standard feed-forward neural networks. The antecedent scoring function incorporates a feature vector encoding speaker and genre information and the distance between the two spans. The mention score and the antecedent score are concatenated to produce the coreference score between the candidate antecedent and the mention span. During training, the model optimizes the marginal log-likelihood of all correct antecedents in the correct clustering $GOLD(i)$ (see Figure 2e for the architecture of the model):

$$
log \prod_{i=1}^{N} \sum_{\hat{y} \in Y(i) \cap GOLD(i)} P(\hat{y})
$$

The input layer of the model consists of a fixed concatenation of 300-dimensional GloVe embeddings[41], 50-dimensional Turian embeddings[42] and 8-dimensional character embeddings. The hidden layers in the LSTMs have 200 dimensions, and the two hidden layers in the feed-forward neural network have 150 dimensions. The speaker, genre information and the span distance and span width are represented as 20-dimensional embeddings. To maintain computing efficiency, the the maximal span width were set to 10, the maximal number of antecedent were set to 250 and the maximal number of sentences in the document was set to 50. The model was trained on the English data from the CoNLL-2012 shared task[43], which contains mainly news articles. The model achieved an average F1 score of 75.8 for a single model on the test set of the English data, outperforming previous neural network models for coreference resolution[44,45] by 1.5 F1. The model was implemented in Tensorflow[46] and the codes are freely available. We took the pretrained English model to generate the softmax of the coreference scores for all the pronouns in the English fMRI and MEG stimuli. We then trained the model on the Chinese data from the CoNLL-2012 shared task[43]. We removed the Turain embeddings which are not available for Chinese, and used only the 300-dimensional word2vec embeddings for Chinese[47] trained on Baidu Encyclopedia. The model achieved an average F1 score of 63.1 for a single model on the test set of the Chinese data from the CoNLL-2012 shared task. We then took the trained Chinese model to generate the softmax of the coreference scores for all the pronouns in the Chinese fMRI stimulus.

Model performance. To evaluate the performance of the four models applied to the three texts, we first computed the model predictions. For the Hobbs model, we computed the Hobbs distance for each of the third person pronouns. Hobbs distances of 1, 2 and 3 indicates correct prediction of the model within three proposals. For the Centering model, we computed the transition type for each entity including the pronoun in the current sentence, and ranks the pronoun's transition type among the transition types for all the entities. For the ACT-R model, we calculated the activation levels for the preceding 20 entities for each third person pronoun. We then ranked the potential mentions according to their ACT-R activation levels; for the NeuralCoref model, we computed the softmax of coreference scores for all the entities from the current sentence containing the pronoun to five sentences preceding the sentence containing the correct antecedent. We then ranked the potential mentions according to the coreference scores. The Centering, ACT-R and the NeuralCoref model are considered correct if the correct antecedent is ranked within the top 3 of the list. This is the SUCCESS@N metric ($N = \{1, 2, 3\}$)[13]. SUCCESS@N is the proportion of instances where the gold answer—the unit label—occurs within a system's first N choices. SUCCESS@1 is standard accuracy. The SUCCESS@N metric allows some degree of ambiguity in selecting the the referents, which parallels human performance during pronoun resolution.

Experiment procedures. fMRI experiments: After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner. Auditory stimuli were delivered through MRI-safe, high-fidelity headphones (English: Confon HP-VS01, MR Confon, Magdeburg, Germany; Chinese: Ear Bud Headset, Resonance Technology, Inc, California, USA) inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. An experimenter increased the sound volume stepwise until the participants could hear clearly. The English and Chinese audiobooks lasted for 94 and 99 minutes, respectively. They were both divided into nine sections, each lasted for about ten minutes. Participants listened passively to the nine sections and completed four quiz questions after each section (36 questions in total). These questions were used to confirm their comprehension and were viewed by the participants via a mirror attached to the head coil and they answered through a button box. The entire session, including preparation time and practice, lasted for around 2.5 hours.

MEG experiment: After giving their informed consent, each participant's head shape was digitized using a Polhemus dual source handheld FastSCAN laser scanner (Polhemus, VT, USA). Participants then completed the experiment while lying supine in a dimly lit, magnetically shielded room. MEG data were recorded continuously using a whole-head 208 channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan). Auditory stimuli were delivered through MEG-safe, high-fidelity earphones inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. An experimenter increased the sound volume stepwise until the participants could hear clearly. The audio lasted for about 12 minutes. Participants listened passively to the audio and completed four picture-matching task. This task was used to confirm their comprehension and were completed by the participants outside the MEG scanner. The entire session lasted for around 30 minutes. All presentation scripts were written in PsychoPy2[48].

Data acquisition and preprocessing. fMRI data: Both English and Chinese brain imaging data were acquired with a 3T MRI GE Discovery MR750 scanner with a 32-channel head coil. Anatomical scans were acquired using a T1-weighted volumetric Magnetization Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence. Blood-oxygen-level-dependent (BOLD) functional scans were acquired using a multi-echo planar imaging (ME-EPI) sequence with online reconstruction (TR=2000 ms; TEs=12.8, 27.5, 43 ms; FA=77°; matrix size=72 x 72; FOV=240.0 mm x 240.0 mm; 2 x image acceleration; 33 axial slices, voxel size=3.75 x 3.75 x 3.8 mm). Cushions and clamps were used to minimize head movement during scanning.

All fMRI data were preprocessed using AFNI version 16[49]. The first 4 volumes in each run were excluded from analyses to allow for T1-equilibration effects. Multi-echo independent components analysis (ME-ICA)[50] were used to denoise data for motion, physiology and scanner artifacts. Images were then spatially normalized to the standard space of the Montreal Neurological Institute (MNI) atlas, yielding a volumetric time series resampled at 2 mm cubic voxels.

MEG data: MEG data were recorded continuously at a sampling rate of 1000 Hz with an online bandpass filter of 0.1-200 Hz. The raw data were first noise reduced via the Continuously Adjusted Least-Squares Method[51] and low-pass filtered at 40 Hz. Independent component analysis (ICA) was then applied to remove artifacts such as eye blinks, heart beats, movements, and well-characterized external noise sources. The MEG data were then segmented into 500 ms epochs at the onset of each word in the stimulus. No baseline correction was applied. Epochs containing amplitudes greater than an absolute threshold of 2000 fT were automatically removed.

Cortically constrained minimum-norm estimates[52] were computed for each epoch for each participant. To perform source localization, the location of the participant's head was coregistered with respect to the sensor array in the MEG helmet using FreeSurfer's[53] "fsaverage" brain, which involved first rotation and translation and then scaling the average brain to match the size of the head scan. A source space of 2562 source points per hemisphere was generated on the cortical surface for each participant. The Boundary Element Model (BEM) was employed to compute a forward solution, explaining the contribution of activity at each source to the magnetic flux at the sensors. Channel-noise covariance was estimated based on the whole epoch. The inverse solution was computed from the forward solution and all the epochs. To lift the restriction on the orientation of the dipoles, the inverse solution was computed with "free" orientation, meaning that the inverse operator places three orthogonal dipoles at each location defined by the source space. When computing the source estimate, only activity from the dipoles perpendicular to the cortex were included. The same inverse operator was applied to each single trial to yield the dynamic statistical parameter maps (dSPM) units54 using an SNR value of 3. All data preprocessing steps were performed using MNE-python (v.0.19.2)[54].

Localise brain regions for third person pronoun processing. fMRI data: A GLM analysis was conducted to localize the active brain regions during third person pronoun processing in both English and Chinese. We first aligned the word boundaries in the English and Chinese fMRI stimuli with the transcripts (see Figure 2a,b for an example). We then modeled the timecourse of each voxel's BOLD signals for each of the nine sections by a binary third person pronoun regressor, time-locked at the offset of each third person pronoun in the audiobook. We included four control variables: the root mean square intensity (`RMS intensity`) for every 10 ms of each audio section, the `f0` of each audio section extracted using the Voicebox toolbox[55], the binary regressor time-locked to the offset of each word in the audio (`word rate`), and the unigram frequency of each word (`frequency`), estimated using the Google ngrams and the SUBTLEX corpora for English[56] and Chinese[57]. These regressors were convolved with SPM12's[58] canonical HRF function and matched the scan numbers of each section.

At the group level, the contrast image for third person pronouns for both English and Chinese were examined by a factorial design matrix. An 8 mm full-width at half-maximum (FWHM) Gaussian smoothing kernel was applied on the contrast images from the first-level analysis to counteract inter-subject anatomical variation. The statistical threshold was set at $p \leq 0.05$ FWE, with an adequate cluster size greater than 20 voxels.

MEG data: A similar two-stage regression analyses was conducted to find the significant spatiotemporal clusters that are correlated with third person pronoun processing. We applied the same regression model to each for each participant's single-trial source estimates for each source at each timepoint of the whole 500 ms time window. This resulted in a β coefficient for each variable at each source and each timepoint for each subject. The source estimates were resampled to 100 Hz.

At the second stage, we performed a one-sample t-test on the distribution of β value for the binary third person pronoun regressor across subjects, again at each source and each timepoint, to test if their values were significantly different from zero. The t-tests are one-tailed such that t-values with the same polarity are clustered together as separate regions. Clusters were then formed based on the t-values that were contiguously significant through time and space, at a level of $p < .05$. Only clusters that contained a minimum of 10 sources and spanned at least 10 ms were entered into a cluster-based spatiotemporal permutation test[19]. This involved randomly shuffling 0 and the β coefficient for each participant, repeating the mass univariate one-sample t-test and cluster formation within the 0-500 ms analysis window. This procedure was performed 10,000 times, resulting in a distribution of 10,000 cluster-level statistics. Each of the observed clusters was subsequently assigned a p-value based on the proportion of random partitions that resulted in a larger test statistic than the observed one. The regression analyses were performed with MNE-python (v.0.19.2)[54] and Eelbrain (v.0.25.2)[59].

RSA within the fROIs. fMRI data: The significant LMTG and RMTG clusters from the GLM analyses were used as the fROIs to compare the four models's relatedness to the brain data. The LMTG fROI contains 294 voxels and the RMTG fROI contains 40 voxels (see **??**c. For each subject in each group, we first extracted all the brain scans after the occurrence of each pronoun by aligning them with the offset of each pronoun in the audiobook. We added 5 seconds to the offset to capture the peak of the hemodynamic response function. The resulting English dataset contains 588 fMRI scans and the Chinese dataset contains 493 fMRI scans. We then regressed out the effects of intensity, f0, word rate and word frequency by subtracting from the brain data the four regressors multiplied by their beta values derived from the GLM analyses. Next, we calculated the RDMs between the brain activity patterns using 1 minus the Pearson's correlation, computed across voxels.

The complexity metrics derived from the Hobbs, Centering, ACT-R and NeuralCoref model for each pronoun in the English and Chinese fMRI stimuli were used to construct the model RDMs, computed as the euclidean distance between each metric value. If multiple pronouns occur within one fMRI scan, the metrics were summed up. The English model RDMs are 588×588 matrices and the Chinese model RDMs are 493×493 matrices. Figure 5a,c show the four model RDMs, each separately rank-transformed and scaled into [0,1].

To compare the four models' ability to explain the fMRI data RDM within the two fROIs, we calculated the Spearman's rank correlation between the data RDMs and each model RDM for each subject in each group. Statistical significance was tested using a one-sided signed-rank test[21] across each subject's correlation values. Pairwise comparison between each model's relatedness to the brain data was also tested using a one-sided signed-rank test. FDR correction was applied for multiple comparisons across fROIs and model pairs.

MEG data: We subset the source estimates of the pronouns within the lateral and medial clusters derived from the regression analyses for third person pronoun processing. At each timepoint, we computed the a MEG data RDM for each pronoun as 1-Pearson correlation between the MEG data patterns.

The complexity metrics derived from the Hobbs, Centering, ACT-R and NeuralCoref model for each pronoun in the MEG stimuli were used to construct the model RDMs, computed as the euclidean distance between each metric value. The model RDM is a 51×51 matrix. Figure 7a shows the four model RDMs, each separately rank-transformed and scaled into [0,1].

We calculated the Spearman's rank correlation between the MEG data RDM and each model RDM at each timepoint for each subject. Statistical significance was tested using a cluster-based permutation *t*-test[19] across each subject's correlation map.

Searchlight RSA. fMRI data: Searchlight RSA was carried out with a spherical cluster (radius=8 mm) for each voxel within a bilateral mask. The mask covered anatomical regions including the superior and middle temporal gyrus (STG, MTG), the inferior frontal gyrus (IFG), the angular gyrus (AG) and the Precuneous cortex(PC) based on the Harvard-Oxford cortical atlas[23] (see Figure 5g). These regions have been previously implicated in pronoun processing.

In each iteration of the searchlight, a brain data RDM and four model RDMs were constructed the same way as in the ROI-based RSA, and the Spearman's rank correlation were calculated between the data RDM and each model RDM. After iterating the searchlight across the searchlight mask, we obtained four maps of Spearman's rank correlation per participant, representing how well the representational geometry in different voxels conforms to the four model's predictions. Statistical inference was examined using one-sided *t*-test across subjects and the resulting *rho* maps were thresholded at $p \leq 0.05$ FWE with a cluster size greater than 50 voxels (see Figure 5g,h). Both the ROI-based and the searchlight RSAa were performed using the RSA toolbox[60].

MEG data: Similar process was applied to the MEG data. The searchlight covers a hexagonal cortical patch (radius = 20 mm) and extends in time for 20 ms. Statistical inference was examined using a cluster-based permutation *t*-test. Clusters were then formed based on the *t*-values that were contiguously significant through time and space, at a level of $p < .05$. Only clusters that contained a minimum of 10 sources and spanned at least 10 ms were entered into a cluster-based spatiotemporal permutation test. The analysis time window is the whole 0-500 ms. All the RSA analyses for the MEG data were conducted using the python men-rsa package[61] and the permutation tests were performed with Eelbrain (v.0.25.2)[59].

Whole-brain GLM analyses fMRI data: To supplement the RSA methods, we also conducted GLM analyses with the metrics derived from the Hobbs, Centering, ACT-R and NeuralCoref model as regressors. Same with the GLM analysis for localizing third person pronouns, we modeled the timecourse of each voxel's BOLD signals for each of the nine sections by the model regressors, time-locked at the offset of each third person pronoun in the audiobook. Under the assumption that increased hemodynamic response of a brain region indicates increased effort related to the task, we took the negative of the ACT-R activation level and the NeuralCoref score to represent the processing difficulty of the pronouns predicted by the model. The control variables include `RMS intensity`, `f0`, `word rate`), `frequency`, and the binary third person pronoun regressor(`pronoun3rd`. These regressors were convolved with SPM12's[58] canonical HRF function.

At the group level, the contrast image for each model regressor was examined by a factorial design matrix. An 8 mm FWHM Gaussian smoothing kernel was applied on the contrast images from the first-level analysis to counteract inter-subject anatomical variation. The statistical threshold was set at $p \leq 0.05\ FWE$, with an adequate cluster size greater than 20 voxels. The GLM analyses was performed separately for both English and Chinese using SPM12[58].

Acknowledgements

This work is supported by the National Science Foundation under Grant No.1607441 and Grant No.1923144. The Chinese data collection is supported by the Jeffrey Sean Lehman Fund for Scholarly Exchange with China at Cornell University. We thank the Jiangsu Key Laboratory of Language and Cognitive Neuroscience at Jiangsu Normal University for help with data collection. We thank Maria Julieta Guzman and Samantha Wray for segmenting the MEG audio stimuli. We thank Shaonan Wang for helpful discussion with the neural coreference model. We thank the High Performance Computing resources at NYUAD for computing support.

Author contributions

J.H. designed the fMRI study. J.L. and L.P. designed the MEG study. J.L., J.H. and W.L collected the English fMRI data. J.L. collected the Chinese fMRI data and the MEG data. J.L. analyzed data. All authors wrote the paper.

Competing interests

The authors declare no competing interests.

References

1. Hobbs, J. in *Readings in natural language processing* (Morgan Kaufman Publishers, Inc., Los Altos, California, USA., 1977).
2. Chomsky, N. *Lectures on government and binding* (Foris, Dordrecht, Holland, 1981).
3. Brennan, S., Friedman, M. W. & Pollard, C. J. *A centering approach to pronouns* in *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, USA, 1987), 155–162.
4. Grosz, B. J., Weinstein, S. & Joshi, A. K. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* **21,** 203–225 (1995).
5. Van Rij, J., van Rijn, H. & Hendriks, P. How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science* **5,** 564–580 (2013).
6. Lee, K., He, L., Lewis, M. & Zettlemoyer, L. *End-to-end neural coreference resolution* in *Proceedings of the 2017 conference on empirical methods in natural language processing* (Association for Computational Linguistics, Copenhagen, Denmark, 2017), 188–197.
7. Lee, K., He, L. & Zettlemoyer, L. *Higher-order coreference resolution with coarse-to-fine inference* in *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 2 (short papers)* (Association for Computational Linguistics, New Orleans, Louisiana, 2018), 687–692.
8. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2,** 4 (2008).
9. Brennan, J. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass* **10,** 299–313 (2016).
10. Anderson, J. R. Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science* **29,** 313–341 (2005).
11. Ge, N., Hale, J. & Charniak, E. *A statistical approach to anaphora resolution* in *Proceedings of the sixth workshop on very large corpora* **71** (1998), 76.
12. Ariel, M. *Accessing noun-phrase antecedents* (Routledge, London, UK, 1990).
13. Kolhatkar, V. & Hirst, G. *Resolving shell nouns* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2014), 499–510.
14. Hammer, A., Goebel, R., Schwarzbach, J., Münte, T. F. & Jansma, B. M. When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research* **1146,** 185–198 (2007).
15. Hammer, A., Jansma, B. M., Tempelmann, C. & Münte, T. F. Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology* **2,** 1–9 (2011).
16. Matchin, W., Sprouse, J. & Hickok, G. A structural distance effect for backward anaphora in Broca's area: An fMRI study. *Brain and Language* **138,** 1–11 (2014).
17. Brodbeck, C., Gwilliams, L. & Pylkkänen, L. Language in Context: MEG evidence for modality-general and -specific responses to reference resolution. *eNeuro* **3,** e0145–16.2016 1–16 (2016).

18. Brodbeck, C. & Pylkkänen, L. Language in context: Characterizing the comprehension of referential expressions with MEG. *NeuroImage* **147,** 447–460 (2017).

19. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Meth.* **164,** 177–190 (2007).

20. Miceli, G. *et al.* The neural correlates of grammatical gender: An fMRI investigation. *Journal of Cognitive Neuroscience* **14,** 618–628 (2002).

21. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1,** 80–83 (1945).

22. Wehbe, L. *et al.* Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* **9,** e112575 (2014).

23. Desikan, R. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31,** 968–980 (2006).

24. Brennan, J. *et al.* Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language* **120,** 163–173 (2012).

25. Brennan, J., Stabler, E., Van Wagenen, S., Luh, W. & Hale, J. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language* **157-158,** 81–94 (2016).

26. Bhattasali, S. *et al.* Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience,* 2327–3801 (2018).

27. Lambon Ralph, M. A., Sage, K. & Roberts, J. Classical anomia: A neuropsychological perspective on speech production. *Neuropsychologia* **38,** 186–202 (2000).

28. Miceli, G., Giustolisi, L. & Caramazza, A. The interaction of lexical and non-lexical processing mechanism: Evidence from anomia. *Cortex* **27,** 57–80 (1991).

29. Oldfield, R. C. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9,** 97–113 (1971).

30. 小王子网站 `http://www.xiaowangzi.org/` (2020).

31. *Scishow Kids* `https://www.youtube.com/user/scishowkids` (2020).

32. Finkel, J. R., Grenager, T. & Manning, C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* 363–370 (2005).

33. Stenetorp, P. *et al. BRAT: a web-based tool for NLP-assisted text annotation* in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2012), 102–107.

34. *The Penn Phonetics Lab Forced Aligner* `https://babel.ling.upenn.edu/phonetics/old_website_2015/p2fa/index.html` (2009).

35. *FAVE (Forced Alignment and Vowel Extraction) Suite Version 1.1.3* `https://www.research.ed.ac.uk/portal/en/publications/fave-forced-alignment-and-vowel-extraction-suite-version-113(bbc2046d-6768-47c5-b574-2987895b0307).html` (2014).

36. De Marneffe, M., MacCartney, B. & Manning, C. *Generating typed dependency parses from phrase structure parses* in *LREC 2006* (2006).

37. Chang, P.-C., Tseng, H., Jurafsky, D. & D., M. C. *Discriminative reordering with Chinese grammatical relations features* in *Proceedings of the third workshop on syntax and structure in statistical translation* (2009).

38. Anderson, J. R. *How can the human mind occur in the physical universe?* (Oxford University Press, Oxford, 2007).

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

40. Bahdanau, D., Cho, K. & Bengio, Y. *Neural machine translation by jointly learning to align and translate* 2016. arXiv: 1409.0473.

41. Pennington, J., Socher, R. & Manning, C. *GloVe: Global vectors for word representation* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Doha, Qatar, 2014), 1532–1543.

42. Turian, J., Ratinov, L.-A. & Bengio, Y. *Word representations: A simple and general method for semi-supervised learning* in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Uppsala, Sweden, 2010), 384–394.

43. Pradhan, S., Moschitti, A., Xue, N., Uryupina, O. & Zhang, Y. *CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes* in *Proceedings of the Sixteenth Conference on Computational Natural Language Learning* (Association for Computational Linguistics, 2012), 1–40.

44. Clark, K. & Manning, C. D. *Improving coreference resolution by learning entity-Level distributed representations* in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* **1** (Association for Computational Linguistics, Berlin, Germany, 2016), 643–653.

45. Clark, K. & Manning, C. D. *Deep reinforcement learning for mention-ranking coreference models* in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2016), 2256–2262.

46. Abadi, M. *et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems* 2016. arXiv: 1603.04467.

47. Li, S. *et al. Analogical Reasoning on Chinese Morphological and Semantic Relations* in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Association for Computational Linguistics, Melbourne, Australia, 2018), 138–143.

48. Peirce, J. W. PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods* **162,** 8–13 (2007).

49. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal* **29,** 162–173 (1996).

50. Kundu, P., Inati, S. J., Evans, J. W., Luh, W.-M. & Bandettini, P. A. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage* **60,** 1759–1770 (2012).

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

52. Hämäläinen, M. S. & Ilmoniemi, R. J. Interpreting magnetic fields of the brain: Minimum norm estimates. *Med. Biol. Eng. Comput.* **32,** 35–42 (1994).

53. *Freesurfer* `http://surfer.nmr.mgh.harvard.edu/` (2020).

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

55. *VOICEBOX: Speech processing toolbox for MATLAB* `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html` (2020).

56. Brysbaert, M. & New, B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* **41,** 977–990 (2009).

57. Cai, Q. & Brysbaert, M. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *Plos ONE* **5,** e10729 (2010).

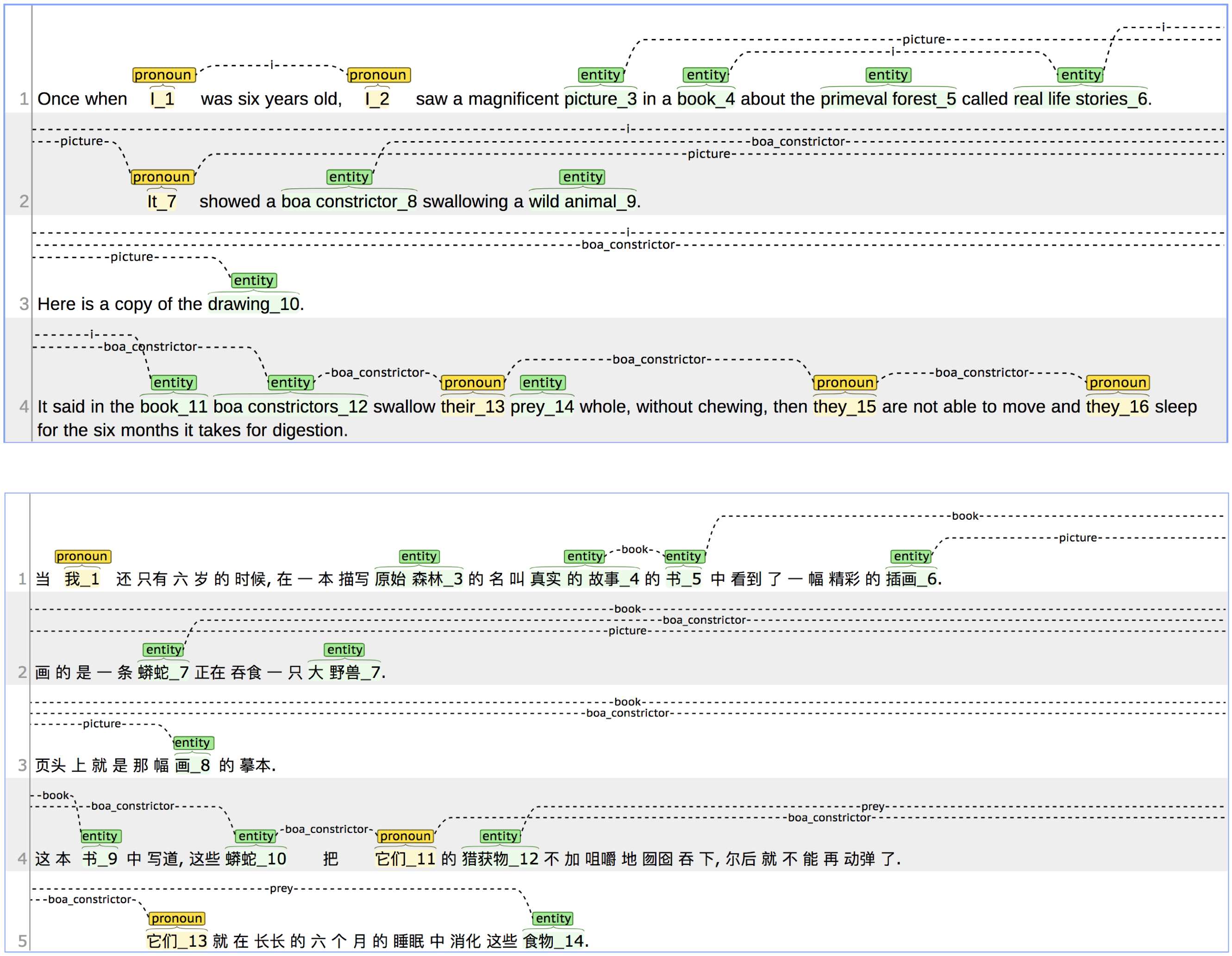
58. Penny, W., Friston, K., Ashburner, J., Kiebel, S. & Nichols, T. *Statistical parametric mapping: The analysis of functional brain images* (Academic Press., 2011).

59. Brodbeck, C. & et al. *Eelbrain-v0.32.* DOI:10.5281/zenodo.3923991 (2020).

60. Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS Computational Biology* **10,** e1003553 (2014).

61. *rsa-mne* `https://github.com/wmvanvliet/mne-rsa` (2020).

Supplementary Information



1 Once when I_1 was six years old, I_2 saw a magnificent picture_3 in a book_4 about the primeval forest_5 called real life stories_6.

2 It_7 showed a boa constrictor_8 swallowing a wild animal_9.

3 Here is a copy of the drawing_10.

4 It said in the book_11 boa constrictors_12 swallow their_13 prey_14 whole, without chewing, then they_15 are not able to move and they_16 sleep for the six months it takes for digestion.

1 当 我_1 还 只有 六 岁 的 时候, 在 一 本 描写 原始 森林_3 的 名 叫 真实 的 故事_4 的 书_5 中 看到 了 一 幅 精彩 的 插画_6.

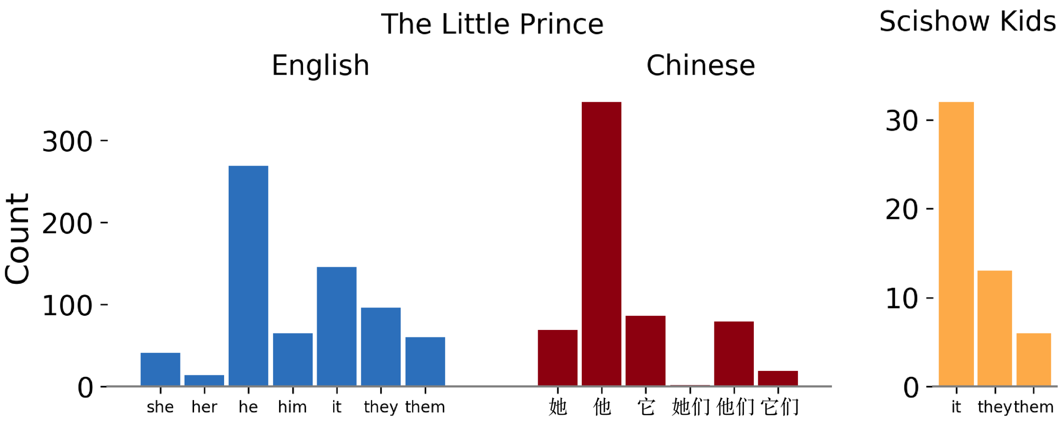
2 画 的 是 一 条 蟒蛇_7 正在 吞食 一 只 大 野兽_7.

3 页头 上 就 是 那 幅 画_8 的 摹本.

4 这 本 书_9 中 写道, 这些 蟒蛇_10 把 它们_11 的 猎获物_12 不 加 咀嚼 地 囫囵 吞 下, 尔后 就 不 能 再 动弹 了.

5 它们_13 就 在 长长 的 六 个 月 的 睡眠 中 消化 这些 食物_14.

Supplementary Figure 1: Sample annotations of pronouns and non-pronoun mentions in *The Little Prince* in English and Chinese, visualized using the annotation tool brat[33].



Supplementary Figure 2: Counts for third person pronouns in the English and Chinese *The Little Prince and the SciShow Kids text.*

	Model	Cluster	MNI coordinates	*k*-size	*t*-value	*p*-value
English	Hobbs	RMTG	60,-22,0	428	6.37	<.0001
	Centering	RSFG	24,16,60	74	5.97	<.0001
		PC	-4,-44,52	60	5.33	0.005
	ACT-R	LIFG	-48, 40, -6	4400	9.99	<.0001
		LMTG	-52, -44, -6	2451	8.87	<.0001
		RMTG	60,-8,0	547	6.88	<.0001
	NeuralCoref	LSTG	-54,-20,-2	1273	7.4	<.0001
		RMTG	48,-26,-8	305	6.18	<.0001
		LAG	-54,-48,22	120	5.57	0.002
		LSFG	-8,48,38	134	5.34	0.005
Chinese	Hobbs	LSMA	-32,-26,28	109	5.61	0.002
	ACT-R	LAG	-52, -64, 26	517	7.92	<.0001
		LMTG	-58,0,-18	281	7.54	<.0001
		LSFG	-10,24,52	426	7.36	<.0001
		LIFG	-52,28,18	72	5.04	0.001
	NeuralCoref	LSTG	-66,-40,10	144	7.33	<.0001
English > Chinese	Hobbs	RMTG	60,-22,0	201	6.03	<.0001
	Centering	PC	2,-72,46	68	5.14	0.01

Supplementary Table 1: MNI coordinates, cluster size and their peak level statistics for the significant clusters derived by the GLM analyses