Cortical processing of reference in language revealed by computational models

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

Shaonan Wang
Chinese Academy of Sciences
Beijing, China

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, USA7

Abstract

Human language processing involves not only combining word meanings in accordance with semantic and syntactic constraints, but also figuring out who and what is being referred to. Here we present a first study towards a mechanistic understanding of the neural basis for referential processing. Using both functional MRI and magnetoencephalography (MEG), we identified a consistent increase of activity in the left middle temporal gyrus for pronoun processing during naturalistic listening for both English and Chinese speakers. We then adopted a "reverse-engineering" approach to examine the cognitive processes underlying pronoun resolution. We evaluated the neural fit of three symbolic models that each formalizes a different strand of explanation for pronoun resolution in the cognitive and linguistic literature, as well as two deep neural network models with either an LSTM or a Transformer architecture. Our results favor the memory-based symbolic model, suggesting a domain-general mechanism of pronoun resolution that resembles memory retrieval.

Introduction

Word meaning involves not only the general, dictionary-like meaning (sense) but also the specific entity it denotes (reference). The classic example used by Frege (1892) is "the morning star": "the morning star" and "the evening star" differ in their sense, but they have the same reference: the Venus. Reference allows us to relate our ideas of things to entities in the world. One example of this is using pronouns such as "she" and "it" to refer to a person and an object in the context. Pronouns are ubiquitous in human language, yet despite decades of research on pronoun resolution in linguistics and cognitive science, the neural basis for pronoun processing remains largely uncharacterised, and it is absent from recent neurobiological models of language comprehension (e.g., Friederici, 2011; Hagoort & Indefrey, 2014). Prior neuroimaging studies on referential processing have reported a number of regions, including the medial parietal lobe (van Berkum et al., 2007; Brodbeck & Pylkkänen, 2017; Brodbeck et al., 2016), the lateral parietal region (van Berkum et al., 2007), the inferior frontal region (Hammer et al., 2007; Matchin et al., 2014; Miceli et al., 1991) and the temporal regions (Hammer et al., 2007, 2011; Miceli et al., 1991), yet they involve different experimental manipulations such as contrasting referentially ambiguous pronouns with non-ambiguous pronouns (van Berkum et al., 2007), contrasting gender-incongruent with gender-congruent pronouns (Hammer et al., 2007). It is therefore

unclear whether these manipulations tapped the same cognitive processes.

As for the cognitive processes relevant for pronoun resolution, various accounts have been proposed in linguistics and cognitive science. The classic Binding Theory (Chomsky, 1981) in formal linguistics suggests that pronoun resolution is constrained by syntactic constraints, such 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 Centering theory (Grosz et al., 1995) in pycholinguisitcs emphasizes the role of discourse coherence when interpreting pronouns and claims that pronouns refer to the most prominent entity in the discourse context. The memory-based account (Nieuwland & Van Berkum, 2006) suggests that pronoun resolution involves maintaining different references in the working memory, hence is influenced by memory resources. These accounts each emphasizes a different aspect of how we link a pronoun to an antecedent expression during language comprehension, yet it is unknown which account best accounts for the brain mechanism of pronoun resolution.

Here we conducted two analyses to examine both the neural correlates and the cognitive processes for pronoun resolution (see Figure 1). We recorded the blood-oxygen-level-dependent (BOLD) signals while both English and Chinese participants listened to the audiobook of "The Little Prince" for about 100 minutes in the fMRI scanner. This naturalistic setting allows us to examine pronoun processing without invoking ungrammatical constructions. The cross-linguistic comparisons between English and Chinese also provide insights on whether linguistic typology would influence the strategies for pronoun resolution, as Chinese pronouns do not distinguish gender in its spoken form ("she", "he" and "it" are all pronounced as "ta" in Chinese). To further examine the temporal dynamics of pronoun processing, we collected a magnetoencephalography (MEG) dataset while English speakers listened to a 12-minute audio excerpt from the YouTube channel "SciShow Kids". We conducted a general linear model (GLM) analysis to localize the fMRI BOLD signals and the MEG source estimates time-locked at each third person pronouns in the narratives. We identified a consistent increase of activity in the left middle temporal gyrus (MTG) across languages and imaging methods, starting from around 150 ms to 250 ms after the onset of the pronouns.

To understand pronoun processing at the algorithmic level, we adopted the integrative "reverse-engineering" approach that links computational models with brain functions. This approach has been successfully applied to understand the neural mechanisms of primate and

human vision (e.g., Bao et al., 2020; Cadena et al., 2019; Kietzmann et al., 2019), as well as the higher-level cognition of language (e.g., A. Anderson et al., n.d.; J. Brennan et al., 2012, 2016; Jackson et al., 2021; Hale et al., 2018; Martin & Doumas, 2019; Toneva & Wehbe, 2019; Toneva et al., 2021; Schrimpf et al., 2021; Wehbe et al., 2014). Here we compared the fit of five computational models for pronoun processing with the fMRI and MEG data within the LMTG functional regions of interest (fROIs) using representational similarity analysis (Kriegeskorte et al., 2008). Three of the models are knowledge-based symbolic models that formalize the Binding Theory (Hobbs, 1977), the Centering Theory (S. Brennan et al., 1987) and the memory account for pronoun resolution (van Rij et al., 2013), while the other two models are data-driven deep neural networks with either a long short-term memory (LSTM) architecture (ELMo; Lee et al., 2017; Lee et al.,2018) or a Transformer architecture (BERT; Joshi et al., 2019). Both ELMo and BERT have shown high performance in Natural Language Processing and hold great promise for simulating the brain's language system (Toneva et al., 2021; Schrimpf et al., 2021). However, our correlational results showed that the memory-based ACT-R model best explains the neural signatures for both the English and Chinese fMRI data and the MEG data, suggesting a domain-general mechanism for pronoun resolution that resembles memory retrieval.

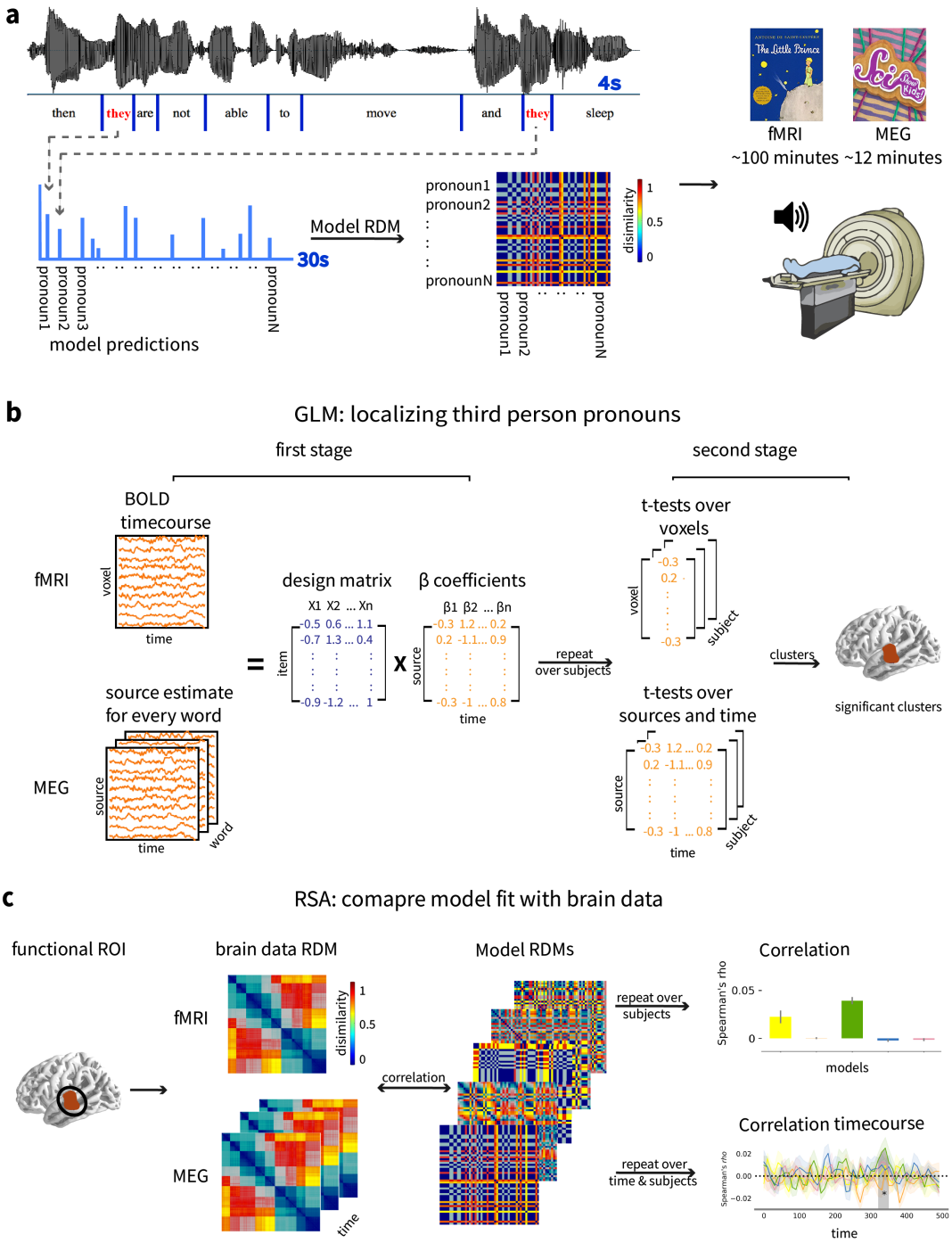
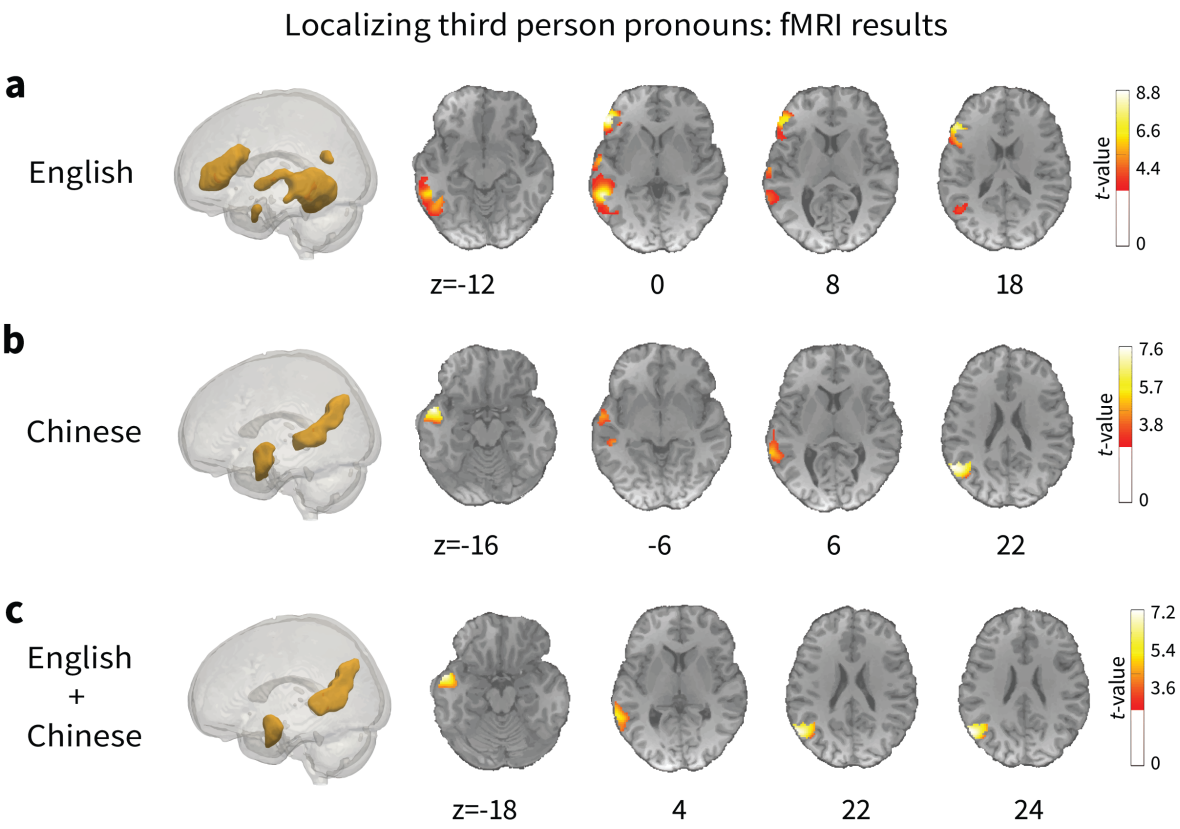


Figure 1: Schematic illustration of the analysis pipeline. **a** Participants listened to the narrative in the fMRI/MEG scanner. Model predictions for all the third person pronouns in the narratives. **b** Localizing third person pronoun processing using GLM. **c** Comparing model relatedness to fMRI/MEG activity pattern within the fROIs derived from the GLM analyses using RSA.

Results

Localizing pronoun processing in space and time. To find brain regions showing increased activation during third person pronoun processing, we modeled the timecourse of each voxel's BOLD signals by a binary third person pronoun regressor, time-locked at the offset of each third person pronoun in the audio. The same model was applied to the MEG source estimate for each word at each timepoint of the whole 500 ms epoch. We focused on third person pronouns as the theoretical accounts for pronoun resolution 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 (Ariel, 1990). We included four control variables: the root mean square intensity (`RMS intensity`) and `f0` for every 10 ms of the audio, and the unigram frequency of each word (`frequency`), estimated using the Google ngrams and the SUBTLEX corpora for English (Brysbaert & New, 2009) and Chinese (Cai & Brysbaert, 2010). Significant clusters for the β coefficient over space and time were identified by a cluster-based spatiotemporal permutation test (Maris & Oostenveld, 2007) with 10,000 permutations (see "GLM analysis" in "Materials and Methods" for details).

The fMRI results for the English data showed significant clusters correlated with the occurrence of third person pronouns in the left inferior frontal gyrus (LIFG; $p = 0.0002$, $k = 1063$), the posterior LMTG ($p = 0.0001$, $k = 2160$) and the anterior LMTG ($p = 0.00288$, $k = 114$; see Figure 2a). For the Chinese data, the GLM analysis revealed similar aLMTG, pLMTG and LIFG clusters ($p = 0.0004$, $k = 581$, $p = 0.0001$, $k = 747$, $p = 0.0404$, $k = 72$, respectively). There is an additional Precuneus activation ($p = 0.0327$, $k = 97$) for the Chinese data (see Figure 2b). The significant regions common to both the English and Chinese data are the aLMTG and the pLMTG clusters ($p = 0.0001$, $k = 1361$ and $p = 0.008$, $k = 492$, respectively). No significant cluster was found for the contrast between the English and Chinese groups.



d

	Cluster	MNI coordinates	*k*-size	*t*-value	*p*-value
English	LIFG	-52,36,0	1063	8.76	0.0002
	pLMTG	-56,-44,-2	2160	8.43	0.0001
	aLMTG	-54,2,-30	114	4.15	0.0288
Chinese	aLMTG	-58,2,-18	581	7.56	0.0004
	pLMTG	-56,-58,22	1554	7.47	0.0001
	Precuneus	-6,-58,34	97	4.63	0.0327
	LIFG	-52,24,16	72	5.04	0.0404
English + Chinese	pLMTG	-58,-62,24	1361	7.2	0.0001
	aLMTG	-58,2,-18	492	6.92	0.008

Figure 2: GLM results of the fMRI data for third person pronoun processing in English and Chinese. **a** Significant clusters of the English fMRI data. **b** Significant clusters of the Chinese fMRI data. **c** Significant clusters of the English and Chinese fMRI data. **d** MNI coordinates and their statistics, thresholded at $p<.05$ *FWE* with a cluster-based permutation test.

The MEG data revealed one significant cluster of 401 sources for third person pronouns from 150 to 250 ms ($p = 0.033$) after the onset of the word. The cluster covered regions including the left posterior temporal lobe (LPTL) and the left medial parietal lobe (Figure 3a,b). Figure 3c,d 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. Compared to other words, third

person pronouns elicited higher activity in the significant time window.

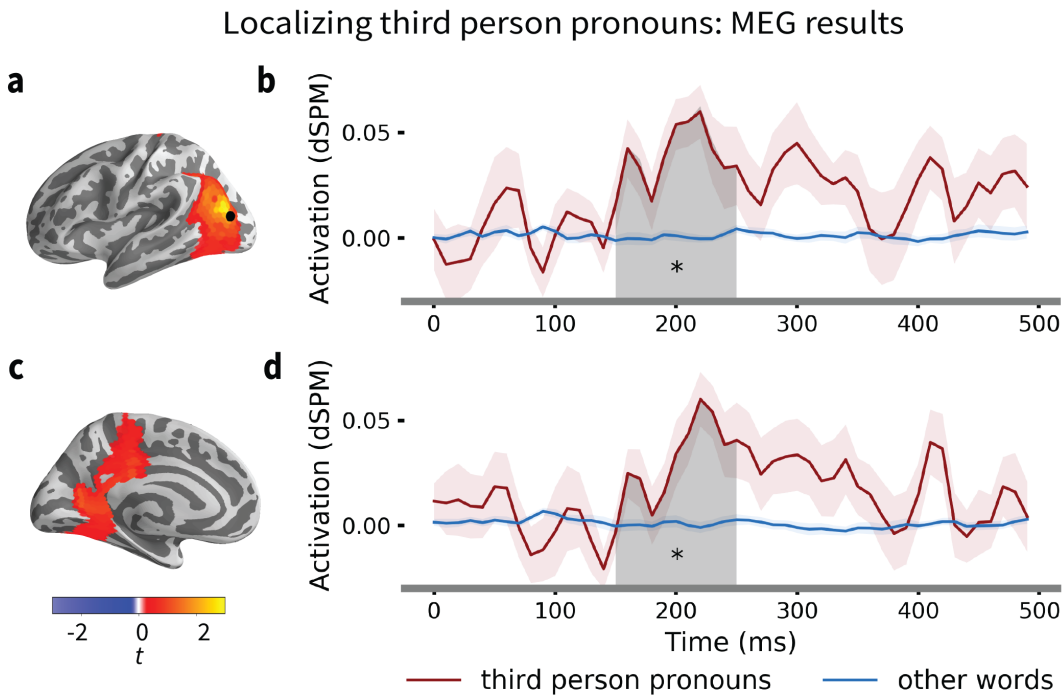


Figure 3: a Lateral part of the significant cluster for third person pronoun processing. **b** Timecourses of responses for third person pronoun and all other words averaged over the lateral cluster. **c** Medial part of the significant cluster for third person pronoun processing. **h** Timecourses of responses for third person pronouns and all other words averaged over the medial cluster. The shaded region indicates the significant time window from 150-250 ms after the word onset ($p = 0.033$).

Both our fMRI and MEG results showed significant LMTG activity, consistent with previous findings (Hammer et al., 2007, 2011; Miceli et al., 2002). Our MEG results showed additional activity in the left medial parietal lobe, which also replicated previous MEG results (Brodbeck et al., 2016; Brodbeck & Pylkkänen, 2017). We extracted the aLMTG and the pLMTG clusters common to both the English and Chinese fMRI data as the functional regions of interests (fROIs) to compare computational models for pronoun resolution with the fMRI data using RSA. Similarly, we split the significant cluster from the MEG results into a lateral part and a medial part and used them as the fROIs to evaluate the models against the MEG data.

Comparing computational models for pronoun resolution. We selected three symbolic models that each formalizes a different strand of explanation for pronoun resolution that has figured in the cognitive and linguistic literature. The syntax-based Hobbs model (Hobbs, 1977) 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 (Chomsky, 1981) 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 discourse-based Centering model (S. Brennan et al., 1987) formalizes the Centering theory (Grosz et al., 1995) 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 (S. Brennan et al., 1987) 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 (van Rij et al., 2013) is specifically intended as a rigorous cognitive model for pronoun resolution. In the modular system of ACT-R (J. R. Anderson, 2005), 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 frequent and more recent chunks are more likely to be retrieved. In addition, spreading activation from other chunks can temporarily boost a chunk's activation. The ACT-R model for pronoun resolution (van Rij et al., 2013) 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 ELMo-Coref model (Lee et al., 2017, 2018) 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". The input to the LSTMs are the ELMo and GloVe embeddings. The BERT-Coref model (Joshi et al., 2019) is an extension of the ELMo-Coref model. It simply replaces the entire LSTM-based encoder with the BERT Transformer (Devlin et al., 2019). Figure 4 shows an illustration of the four models applied to an example English sentence from "The Little Prince". (see "The Hobbs model", "The Centering model", "The ACT-R model", "The ELMo-Coref model" and "The BERT-Coref

model" in "Materials and Methods" for a detailed description of the models).

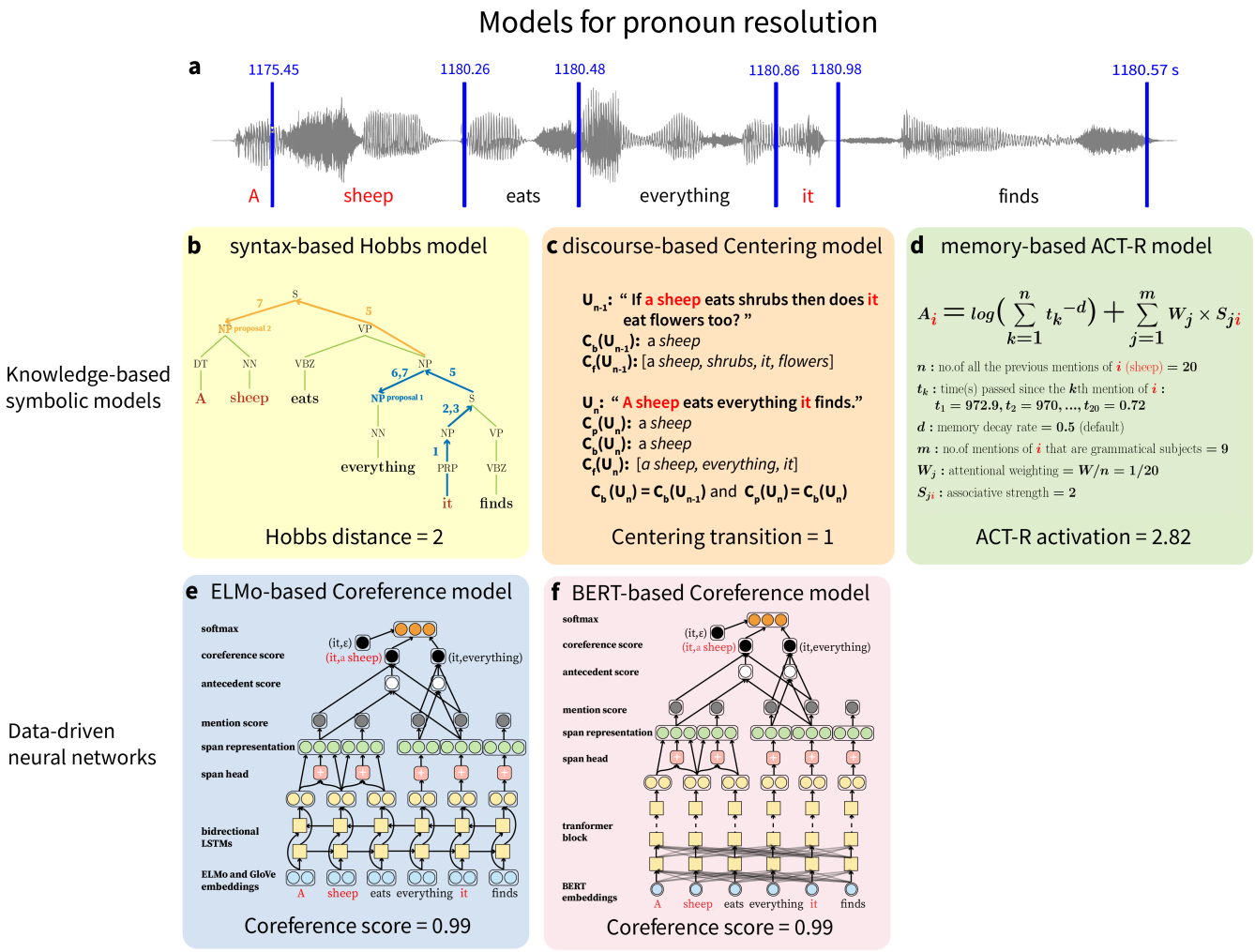


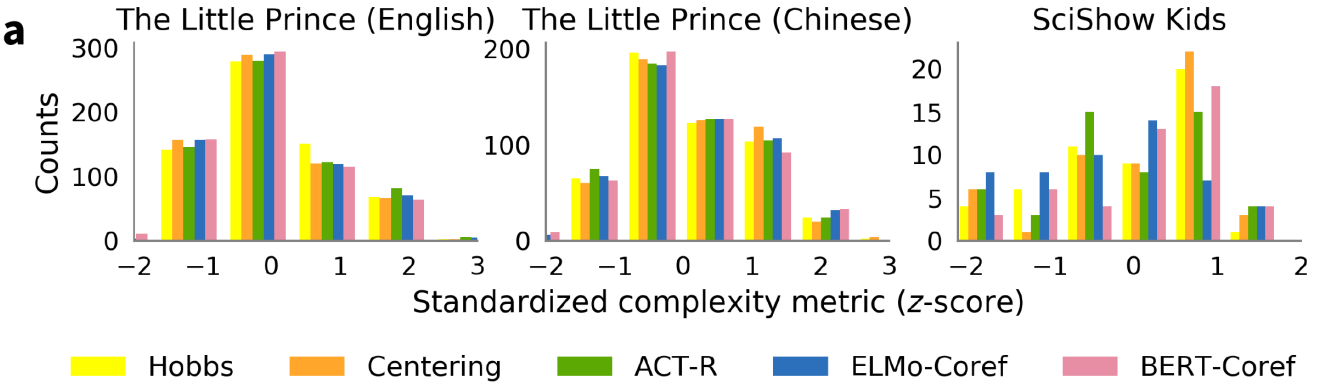
Figure 4: Illustration of the Hobbs (Hobbs, 1977), Centering (S. Brennan et al., 1987), ACT-R (van Rij et al., 2013), ELMo-Coref (Lee et al., 2017, 2018) and BERT-Coref (Joshi et al., 2019) models using an example sentence in the English *The Little Prince*. **a** Waveform of the example sentence from 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 ELMo-Coref model (Lee et al., 2017, 2018). The coreference score is the softmax of the final layer. **f** The architecture of the BERT-Coref model (Joshi et al., 2019). The coreference 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 intervened

by another candidate antecedent on a syntactic tree are hard for the Hobbs model; pronouns not referring to the most prominent entity in the discourse context are hard for the Centering model; pronouns referring to entities that are mentioned only a few times a while back are hard for the ACT-R model, and pronouns surrounded by nouns with similar word meanings are hard for the ELMo-Coref and the BERT-Coref model. To connect properties of the four models to the observed brain data, we define 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" metric (Ge et al., 1998), 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 ELMo-Coref and the BERT-Coref models, we used the negative of coreference score of the antecedent for each pronoun. These complexity metrics allow us to estimate model-derived brain states for comparison against observed brain data.

Figure 5a shows the distribution of the standardized complexity metrics derived by the five models applied to all the third person pronouns in "The Little Prince" in English and Chinese and "SciShow Kids". The distribution of the complexity metrics are very similar for the "The Little Prince" in English and Chinese, where the 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 models. For the MEG stimuli, the complexity metrics are left-skewed with more negative values, suggesting more easier pronouns for the models. However, there are more pronouns with a complexity z-score of 1 than pronouns with mean complexity ($z$score=0), indicating more pronoun with medium difficulty levels in the MEG stimuli.

To ensure that the five models can indeed predict the correct antecedents, we calculated the accuracy of the five 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 (Kolhatkar & Hirst, 2014) where the gold answer occurs within a system's first N choices. All the models performed well with an accuracy well above or near 70% at SUCCESS@3 (see Figure 5b).



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
ELMo-Coref	74.7	80.5	83.6	71.6	77.3	80.3	58.5	72.5	76.5
BERT-Coref	74.2	78.5	81.1	76	80.5	83.4	70.6	74.5	78.4

Figure 5: Comparison of the five models applied to the fMRI and MEG stimuli. **a** Distribution of the standardized model metrics. We took the negative of the ACT-R, ELMo-Coref and BERT-Coref 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, ELMo-Coref and BERT-Coref models based on SUCCESS@N (N=1,2,3) (Kolhatkar & Hirst, 2014), i.e., the proportion of the correct antecedent occurs within the model's first three choices.

Comparing model predictions within the fROIs of the fMRI data. We performed RSA (Kriegeskorte et al., 2008) for each model within the fROIs for third person pronoun processing to compare the model predictions within the significant clusters for third person pronoun resolution in English and Chinese. RSA characterizes a representation in the brain or a 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 used the complexity metrics derived from the five models for each third person pronoun in the fMRI stimuli to construct the model RDMs. The data RDMs were computed as 1 minus Pearson's correlation among the BOLD responses (see "RSA within the fROIs" in "Materials and Methods" for details). Figure 6a,c show the RDMs for the Hobbs, Centering, ACT-R, ELMo-Coref and BERT-Coref model in English and Chinese. Spearman's rank correlation

among the five models is shown in Figure 6b,d.

The RSA results for the English and Chinese data are shown in Figure 6e,f. For the English data, the ACT-R model showed the highest Spearman's rank correlation with both the aLMTG and pLMTG activities, averaged across subjects ($rho = 0.039, p < .0001$ and $rho = 0.042, p < .0001$, respectively). The Hobbs model was also significantly related to both fROI activities ($rho = 0.023, p = 0.0015$ and $rho = 0.028, p < .0001$, respectively). The Centering, ELMo-Coref and BERT-Coref model RDMs were not significantly correlated with either of the fROI activities (Centering: $rho = 0.0004, p = 0.47$ and $rho = 0.0015, p = 0.23$; ELMo-Coref: $rho = -0.0021, p = 0.13$ and $rho = -0.0004, p = 0.57$, BERT-Coref: $rho = -0.0016, p = 0.87$ and $rho = -0.0016, p = 0.87$).

For the Chinese data, the ACT-R model also had the highest mean correlation value with both the aLMTG and pLMTG regions ($rho = 0.07, p < .0001$ and $rho = 0.077, p < .0001$, respectively). The Hobbs model was also significant for both fROIs with a rho value of 0.019 ($p = .002$) and 0.024 ($p < .0001$), respectively. The ELMo-Coref model showed a significant but very low correlation with the aLMTG activity ($rho = 0.003, p = 0.005$) but not the pLMTG activity ($rho = 0.0014, p = 0.66$). The Centering and BERT-Coref models were not significant for either of the fROIs (Centering: $rho = -0.001, p = 0.66$ and $rho = 0.0014, p = 0.7$; BERT-Coref: $rho = -0.0038, p = 0.94$ and $rho = -0.0064, p = 0.99$).

A two-sample t-test between the English and Chinese speakers revealed a significantly higher correlation of the ACT-R models with the Chinese fMRI data in both the aLMTG and pLMTG clusters ($t(82) = 4.6, p < .0001$ and $t(82) = 5, p < .0001$), suggesting a better fit of the ACT-R model for pronoun resolution in Chinese.

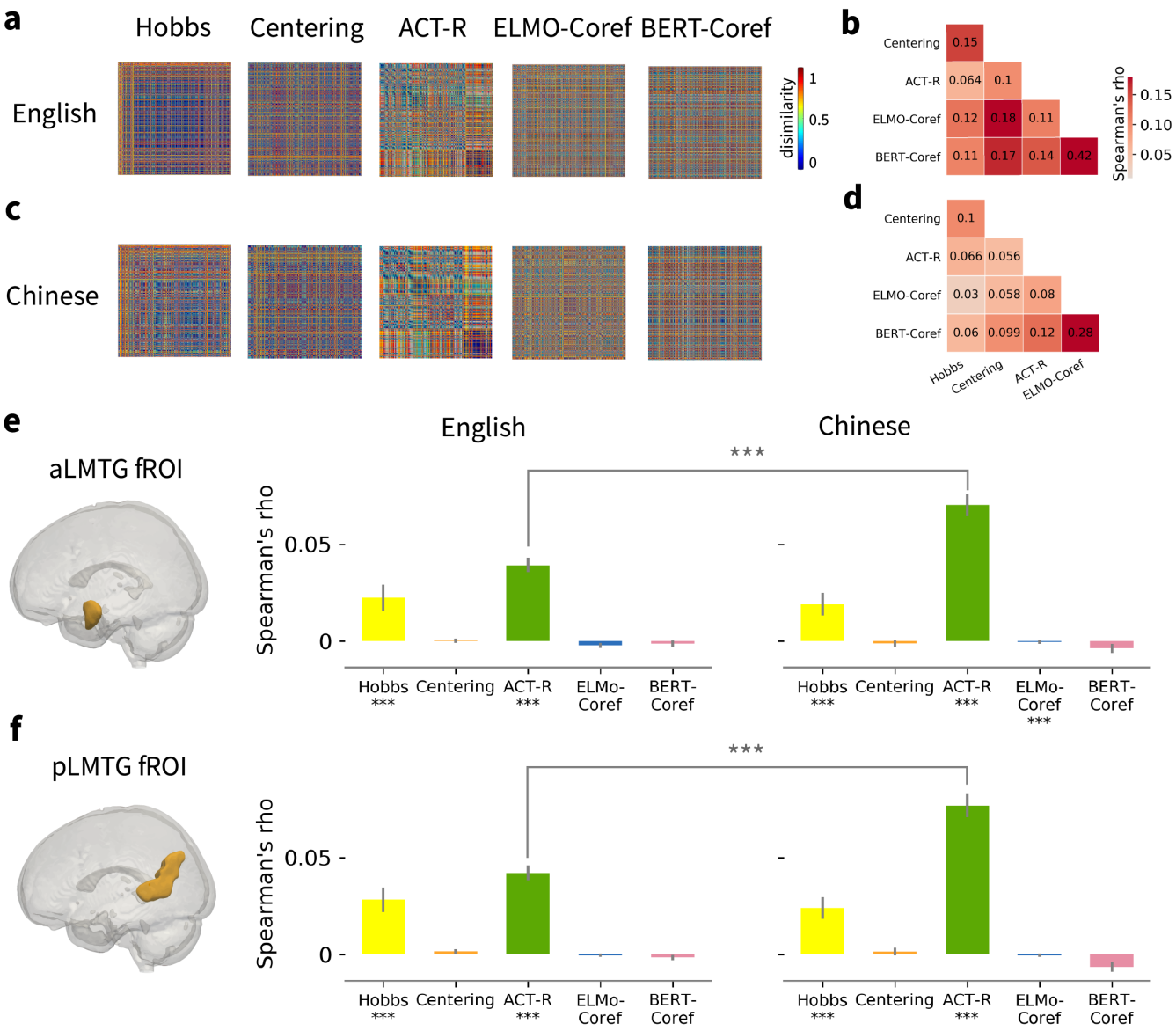


Figure 6: ROI-based RSA results for the fMRI data. **a** The RDMs for the Hobbs, Centering, ACT-R, ELMo-Coref and BERT-Coref model in English. Each RDM was separately rank-transformed and scaled into [0,1]. **b** Spearman's rank correlation matrix for the five model RDMs in English. **c** The model RDMs in Chinese. **d** Spearman's rank correlation matrix for the five model RDMs in Chinese. **e** Relatedness of the five model RDMs to the brain data RDM within the aLMTG fROI, averaged across subjects in each group. **f** Relatedness of the five model RDMs to the fMRI data RDM within the pLMTG fROI, averaged across subjects in each group. * below the model names denotes significant correlation. * above the bars denotes significant group difference. FDR correction was applied for multiple comparisons across fROIs and models. *** $p < .0001$.

Comparing model predictions within the fROIs of the MEG data. For the MEG data, we subset the source estimates of the pronouns within the lateral and medial fROIs derived from the GLM analysis for third person pronoun processing. At each timepoint, we computed the MEG

data RDM for each pronoun as 1 minus Pearson's correlation between the MEG source estimates for all the third person pronouns. The complexity metrics derived from the five models for each pronoun in the MEG stimuli were used to construct the model RDMs, computed as the euclidean distance between each metric value. Figure 7a shows the five model RDMs, each separately rank-transformed and scaled into [0,1]. Figure 7b shows the Spearman's rank correlation matrix for the five model RDMs.

The RSA results showed a significant cluster for the ACT-R model from around 320 to 350 ms after the onset of the pronouns for both the lateral fROI ($p = 0.049$) and the medial fROI ($p = 0.038$). None of the other three models was significantly correlated with the MEG data patterns within either fROI (see Figure 7c-f).

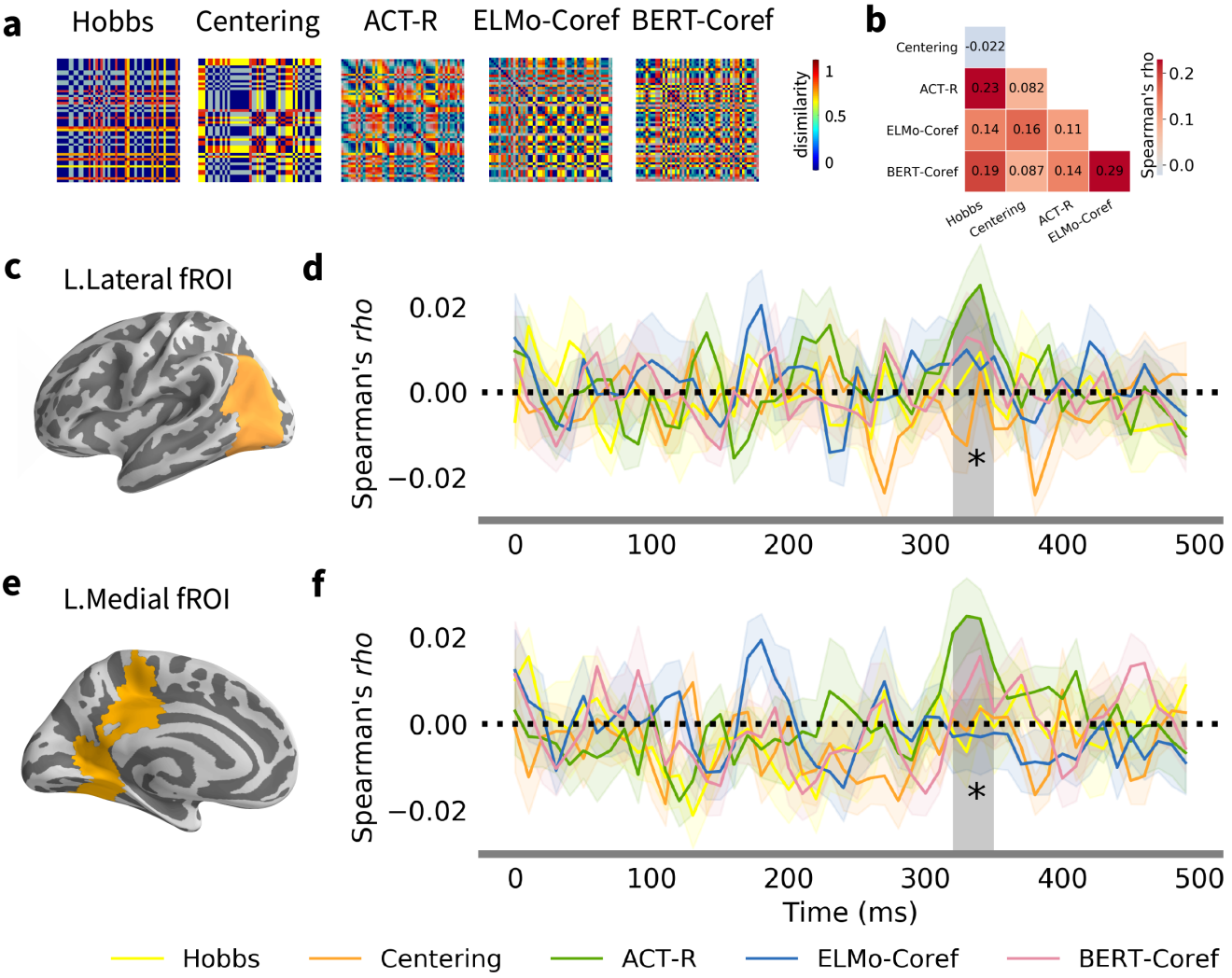


Figure 7: ROI-based RSA results for the MEG data. **a** The RDMs for the Hobbs, Centering, ACT-R, ELMo-Coref and BERT-Coref model metrics for the MEG stimuli. Each RDM was separately rank-transformed and scaled into [0,1]. **b** Spearman's rank correlation matrix for the five model RDMs. **c** fROI in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

Discussion

The aim of the study is to investigate both the neural correlates and the cognitive mechanisms underlying referential processing during naturalistic language comprehension. Our fMRI and MEG results converged on the LMTG for third person pronoun processing across English

and Chinese. The LMTG has been previously associated with biological and syntactic gender processing (Heim et al., 2002; Hammer et al., 2007, 2011; Miceli et al., 2002), and morphological inflection (Longe et al., 2007). For example, Hammer et al. (2007) showed that German sentences with congruent biological and syntactic gender were correlated with increased activation in the LMTG; Miceli et al. (2002) reported the LMTG activity when the subjects were asked whether a written noun has a masculine or a feminine gender. Longe et al. (2007) found greater LMTG activity for plurals and inflected verbs compared to stem words. Both gender and inflections are relevant for pronoun resolution in English since English pronouns distinguish gender and Case (i.e., "she" for subject and "her" for object). However, we found that the LMTG is also implicated for pronoun processing in Chinese, where pronouns do not have gender ("she", "he" and "it" are all pronounced as "ta" in Chinese) or Case inflections. We therefore suggest that the LMTG supports the retrieval of reference in general, consistent with Indefrey & Levelt's (2004) model on language production where the LMTG is the site for lemma selection and retrieval. Furthermore, lesions in the LMTG led to selective difficulty in accessing nouns (Lambon Ralph et al., 2000; Miceli et al., 1991), supporting the LMTG's role in retrieving references.

The MEG data showed an additional recruitment of the left medial parietal region for third person pronoun processing, and the Chinese fMRI data also showed a significant cluster in the left Precuneus. This medial parietal activity has been previously reported when retrieving referentially ambiguous pronouns, as in "Ronald told Frank that he ..." (Nieuwland et al., 2007), or when the sentence contains two referents, as in "Jeremy and Lucy did some work on the house next door." (Boiteau et al., 2014). These findings all suggest that the left medial parietal region may be responsible for tracking multiple referents. Almor et al. (2007) further suggested that the parietal regions might be originally devoted to perceptual organization where it tracks multiple objects in space. Indeed, Brodbeck et al. (2016) used a visual world paradigm in MEG and found the medial parietal activity for successful reference resolution. Therefore, one interpretation of the medial parietal activity in the MEG data and the Chinese fMRI data might be that there are more ambiguous pronouns in the MEG stimuli. Moreover, due to the lack of gender and case information, Chinese speakers may encounter more difficulty in processing referentially ambiguous pronouns.

The MEG data also provide information on the neural activity for pronoun processing with fine temporal resolution. We show that pronoun processing starts from 150 ms till 250 ms

after the onset of the pronoun, consistent with Hauk et al.'s (2012) finding that the brain starts retrieving lexical and semantic information within around 200 ms of word onset in the LMTG. We are aware of only two MEG studies on referential processing in the literature Brodbeck & Pylkkänen (2017); Brodbeck et al. (2016), and our time window is a bit earlier than the 350-500 ms window previously reported. However, the previous MEG studies used a visual word paradigm with multiple candidate referents, which might be more difficult to resolve than the pronouns in our naturalistic listening setting. Moreover, the previous analysis time window was restricted to 200-500 ms after the target stimuli, which is different from the 0-500 ms analysis time window in our study.

To further understand the fine details of the cognitive processes underlying pronoun resolution in the brain, we leveraged computational models for pronoun resolution. Computational models 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 two data-driven neural network models for pronoun resolution against both fMRI and MEG data. Our results all favor the ACT-R model (van Rij et al., 2013): For both the English and Chinese fMRI data, the ACT-R model showed the highest correlation with the BOLD response patterns within the aLMTG and pLMTG fROIs; for the MEG data, the ACT-R model is the only model that showed significant correlation with the source-localized MEG data within the left posterior temporal lobe and the left medial parietal fROIs. Our results therefore support the memory-based account for pronoun resolution (Nieuwland & Van Berkum, 2006), which emphasizes the role of memory resources while retrieving the correct references.

Working memory capacity has long been associated with individual differences in measures of language comprehension in the psycholinguistic literature, and there is a wealth of data suggesting that individuals who perform better on a typical verbal working memory task like the Reading Span task also perform better on both off-line and on-line measures of language comprehension (e.g., Daneman & Carpenter, 1980, 1992). The concept of "memory load" has also been related to complexity in sentence processing. For example, Gibson (1998) claimed that object relative clauses such as "The reporter that the senator attacked disliked the editor." are more difficult to process than subject relative clauses such as "The reporter that attacked the senator disliked the editor.", because the syntactic structure of object relative clauses leads to

higher memory load: The distance between the subject ("the reporter") and its "trace" position is longer in object relative clauses than in subject relative clauses (i.e., the trace position for "the reporter" in the object relative clause is after the verb "attack" whereas the trace in the subject relative clause is before "attack") .

The memory-based account for pronoun resolution employ the a similar concept of memory resource in psycholinguistics, it views pronoun resolution as retrieving the most salient entities from declarative memory. Built using the primitives of the memory module in the cognitive architecture ACT-R (J. R. Anderson, 2005), the ACT-R model 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 ELMo-Coref 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 (J. R. Anderson, 2005). Our correlational results suggest that this algorithm indeed correlates better with the brain's computation states than other non-brain-inspired models.

The two deep neural network models for coreference resolution did not correlate well with the brain data. Deep learning models have led to significant advances in many aspects of natural language processing, and there has been considerable enthusiasm to relate the model architectures with the neural architectures for vision (e.g., Bao et al., 2020; Cadena et al., 2019; Kietzmann et al., 2019) and language (e.g., A. Anderson et al., n.d.; Toneva et al., 2021; Schrimpf et al., 2021). However, many of the deep learning models are not intended to match the human cognitive process. The two neural coreference models here, for example, treated coreference resolution as a clustering task, which is intuitively quite different from how humans comprehend pronouns. Moreover, both the ELMo and BERT architectures are bidirectional with information available from the following discourse, which the human listeners do not have access to during naturalistic listening comprehension. Therefore, although the neural network models performed quite well on the coreference task, they did not match the brain's mechanism of referential analysis.

To sum up, we show that referential processing in human language is primarily localized at the LMTG, and we provide the first evidence that, in natural stories, brain activity for pronoun processing is best explained by the declarative memory module in the ACT-R cognitive architecture. Accordingly, reference resolution in rich contextual settings can be explained via

domain-general, neurobiological mechanisms of memory retrieval in the human brain. Our work strengthens connections between cognitive neuroscience, psycholinguistics and natural language processing. By testing computational models rigorously against neuroimaging data, we also hope to provide insights on designing machines that learn and think like humans. We call for a joint force of cognitive neuroscience and artificial intelligence to explicate the intricate details of the human mind.

Materials and methods

Participants. The English fMRI data were taken from a published study (Bhattasali et al., 2018), collected at the same time with the Chinese data. Participants were 49 young adults (30 females, 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 (Oldfield, 1971). 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 were 35 young adults (15 females, mean age=19.3, SD=1.6) 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 (Oldfield, 1971). 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 females, 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 a 94-minute audiobook 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 99-minute Chinese translation of *The Little Prince* ("小王

子网站", 2020), read by a professional female Chinese broadcaster hired by the experimenter. The MEG stimulus is a 12-minute audio excerpt taken from the YouTube channel "SciShow Kids". It consists of 4 short audios that introducing scientific fun facts to kids. Word boundaries in the audios were identified and aligned to the transcripts using the Forced Alignment and Vowel Extraction (FAVE) and were manually checked by native English and Chinese speakers.

All entities (or mentions; e.g., nouns, pronouns, proper names, etc.) in the audio texts were first identified using the Stanford Named Entity Recognizer (NER) (Finkel et al., 2005). They were then manually checked and linked with their coreferential mentions using the annotation tool brat (Stenetorp et al., 2012). 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 then selected only the third person pronouns (e.g., he, him, she, her, it, they, them) from the mentions as our focus of analyses as the Hobbs and Centering models for pronoun resolution are mainly concerned with third person pronouns. In addition, the cognitive processes may differ for first, second and third person pronoun processing (Ariel, 1990). We removed possessives, reflexives, cleft and extraposition "it", pleonastic "it"and pronouns with sentential antecedents. The final English fMRI stimulus contains 647 third person pronouns, the Chinese fMRI stimulus contains 524 third person pronouns, and the MEG stimulus contains 51 third person pronouns (see Supplemental Figure 2).

The Hobbs model. The Hobbs model for pronoun resolution (Hobbs, 1977) 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. When applied to the Chinese text, the Hobbs model 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. 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 parsed 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.

We use the "Hobbs distance" (Ge et al., 1998) 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 backward from the pronoun, before the correct antecedent NP is found. Figure 4c,d illustrate the Hobbs model for one example sentence in English. 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 (S. Brennan et al., 1987) (also known as the BFP algorithm) formalizes the Centering Theory (Grosz et al., 1995). In the Centering framework, entities that link an utterance to other 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 rank 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 ($C_b(U_n) = C_b(U_{n-1})$), 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 is preferred over RETAINING, which is preferred over SHIFTING. 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 (S. Brennan et al., 1987) 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 SHIFTING 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 ordering of the transitions and select the pair that has the preferred transition types. The rank of the entities in C_f is determined by their grammatical roles, which is parsed using the Stanford dependency

parser for English (de Marneffe et al., 2006) and Chinese (Chang et al., 2009).

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 4c 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 (van Rij et al., 2013) uses the same primitives of the memory module in the cognitive architecture of ACT-R (J. R. Anderson, 2007). The formula for the activation level of 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 (J. R. Anderson, 2007). Different mentions of the entity i are summed 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 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 that are the subjects of their sentences. 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 (de Marneffe et al., 2006; Chang et al., 2009).

The effects of frequency and recency are folded into the calculation of the base activation for antecedent i, such that the more frequent and recent the previous mentions, the higher the base activation. Conversely, if antecedent i has been mentioned only once, or if it was mentioned a long time ago, its activation level will be lower, 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 activation level of an entity in the discourse context is computed based on recency, frequency and the grammatical role of the entity. The entity that has the highest activation level is predicted to be the antecedent of the pronoun. Figure 4d shows how the ACT-R activation level for a pronoun in the English example sentence is calculated.

The ELMo-Coref model. The ELMo-based coreference model (Lee et al., 2017, 2018) is an end-to-end coreference resolution system that predicts all clusters of coreferential mentions given a text document. 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 $START(i)$ and an end index $END(i)$ 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 (Hochreiter & Schmidhuber, 1997):

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

The model then assigns weights to the word vectors to represent the notion of syntactic head using an attention mechanism (Bahdanau et al., 2016):

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

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 an 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 4e 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 the ELMo (Matthew et al., 2018) and GloVe embeddings (Pennington et al., 2014). 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, span distance and span width information are represented as 20-dimensional embeddings. To maintain computing efficiency, the maximal span width was set to 10, the maximal number of antecedent was 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 (Pradhan et al., 2012) and achieved an average F1 score of 75.8. We took the pretrained English model (Lee et al., 2017, 2018) to generate the softmax of the coreference scores for all the pronouns in the English "The Little Prince" and the "SciShow Kids" texts. We then trained the model on the Chinese data from the CoNLL-2012 shared task (Pradhan et al., 2012). We used the pretrained ELMo embeddings for Chinese (Che et al., 2018) and the 300-dimensional word2vec embeddings (Li et al., 2018) 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 "The Little Prince".

The BERT-Coref model. The BERT-based coreference model (Joshi et al., 2019) is an extension of the ELMo-based coreference model (Lee et al., 2017, 2018). It simply replaces the ELMo and GloVe embeddings and the LSTM encoders in the ELMo-Coref model with the BERT Transformers (Devlin et al., 2019). BERT's model architecture is a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017). The BERT-base model consists of 12 identical layers, and each layer contains a multi-head self-attention layer and a fully connected feed-forward network. The attention mechanism maps queries (Q) and a set of key-value (K,V) pairs to an output using a scaled dot-product function, where the queries and keys are vectors of dimension d_k and the values are vectors of dimension d_v:

$$Attention(Q, K, V) = softmax(\frac{QK^t}{\sqrt{d_k}})V$$

BERT-base employed 12 parallel attention layers, or heads, and the output from a single attention function are concatenated and feed to the feed-forward network layer:

$$FFNN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

We used the pretrained BERT-base model for the English data (Joshi et al., 2019) to generate the softmax of the coreference scores for all the pronouns in the English "The Little Prince". We then trained the same model on the Chinese data from the CoNLL-2012 shared task (Pradhan et al., 2012). The model achieved an average F1 score of 65.11 for a single model on the test set of the Chinese data from the CoNLL-2012 shared task. We took the trained Chinese model to generate the softmax of the coreference scores for all the pronouns in the Chinese "The Little Prince".

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. Both the English and Chinese stimuli were divided into 9 sections, and each lasted for about 10 minutes. Participants listened passively to the 9 sections and completed 4 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 tasks. These tasks were 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 (Peirce, 2007).

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). All fMRI data were preprocessed using AFNI version 16 (Cox, 1996). The first 4 volumes in each run were excluded from analyses to allow for T1-equilibration effects. Multi-echo independent components analysis (ME-ICA) (Kundu et al., 2012) 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 (Adachi et al., 2001) 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. The epochs were baseline-corrected using the mean of the whole epoch. Epochs containing amplitudes greater than an absolute threshold of 2000 fT were automatically removed.

Cortically constrained minimum-norm estimates (Hämäläinen & Ilmoniemi, 1994) 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 ("Freesurfer", 2020) "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 activities 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) units using an SNR value of 3. All data preprocessing steps were performed using MNE-python (v.0.19.2) (Gramfort & et al., 2014).

Localize brain regions for third person pronoun processing. For the fMRI data, we 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, 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 (Brysbaert & New, 2009) and Chinese (Cai & Brysbaert, 2010). These regressors were convolved with SPM12's (Penny et al., 2011) canonical HRF function and matched the scan numbers of each section. At the group level, the contrast image for 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. Statistical significance was determined by a cluster-based permutation test (Maris & Oostenveld, 2007), thresholded at $p < .05\ FWE$.

A similar two-stage regression analysis was conducted to find the significant spatiotemporal clusters that are correlated with third person pronoun processing. We applied the same regression model to each participant's single-trial source estimates 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 the β values for the binary third person pronoun regressor across subjects, again at each source and each timepoint. 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 (Maris & Oostenveld, 2007). 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 GLM analysis was performed with MNE-python (v.0.19.2) (Gramfort & et al., 2014) and Eelbrain (v.0.25.2) (Brodbeck & et al., 2020).

RSA within the fROIs. For the fMRI data, the significant aLMTG and pLMTG clusters from the GLM analyses were used as the fROIs to compare the five models' relatedness to the brain data (see Figure 2c). 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. 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, ELMo-Coref and BERT-Coref models 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.

To compare the five 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 determined using a one-sided signed-rank test (Wilcoxon, 1945) across each subject's correlation values. Pairwise comparisons between each model's relatedness to the brain data were also tested using a one-sided signed-rank test. FDR correction was applied for multiple comparisons across fROIs and model pairs.

For the 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 MEG data RDM for each pronoun as 1 minus Pearson's correlation among the MEG data. The complexity metrics derived from the five models 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 RDMs are 51×51 matrices.

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 determined using a cluster-based permutation t-test (Maris & Oostenveld, 2007) with 10,000 permutations across each subject's correlation map. Clusters were formed based 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. FDR correction was applied for multiple comparisons across fROIs and models.

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 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. and S.W

analyzed data. All authors wrote the paper.

Competing interests

The authors declare no competing interests.

References

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

Almor, A., Smith, D. V., Bonilha, L., Fridriksson, J., & Rorden, C. (2007). What is in a name? Spatial brain circuits are used to track discourse references. *Neuroreport*, *18*, 1215-1219.

Anderson, A., Kiela, D., Binder, J., Fernandino, L., Humphries, C., Conant, L., ... Lalor, E. (n.d.). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*, 313-341.

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

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London, UK: Routledge.

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate.*

Bao, P., She, L., McGill, M., & Tsao, D. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*, 103–108.

Bhattasali, S., Fabre, M., Luh, W.-M., Al Saied, H., Constant, M., Pallier, C., ... Hale, J. (2018). Localising memory retrieval and syntactic composition: an fmri study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 2327-3801.

Boiteau, T. W., Bowers, E., Nair, V. A., & Almor, A. (2014). The neural representation of plural discourse entities. *Brain and Language*, *137*, 130-141.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*, 163-173.

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

Brennan, S., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics* (p. 155-162). Stroudsburg, PA, USA: Association for Computational Linguistics.

Brodbeck, C., & et al. (2020). Eelbrain-v0.32. [Computer software manual]. (DOI:10.5281/zenodo.3923991)

Brodbeck, C., Gwilliams, L., & Pylkkänen, L. (2016). Language in context: Meg evidence for modality-general and -specific responses to reference resolution. *eNeuro*, *3*, e0145-16.2016 1–16.

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

Brysbaert, M., & New, B. (2009). 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.

Cadena, S., Denfield, G., Walker, E., Gatys, L., Tolias, A., Bethge, M., & Ecker, A. (2019). Deep convolutional 363 models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, *15*, 1–27.

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

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

Che, W., Liu, Y., Wang, Y., Zheng, B., & Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 55–64).

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, Holland: Foris.

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

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.*, *19*, 450–466.

Daneman, M., & Carpenter, P. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.*, *99*, 122–149.

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

Devlin, J., Ming-Wei, C., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt* (pp. 4171–4186).

Finkel, J. R., Grenager, T., & Manning, C. (2005). 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.

Freesurfer [Computer software manual]. (2020). (`http://surfer.nmr.mgh.harvard.edu/`)

Frege, G. (1892). On sense and nominatum. In A. Martinich (Ed.), *The philosophy of language* (pp. 190–202). New York: Oxford University Press,.

Friederici, A. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *4*, 1357–1392.

Ge, N., Hale, J., & Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora* (Vol. 71, p. 76).

Gibson, E. (1998). Syntactic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gramfort, A., & et al. (2014). Mne software for processing meg and eeg data. *NeuroImage*, *86*, 446–460.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*, 203-225.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, *37*, 347–362.

Hale, J., Dyer, C., Kuncoro, A., & R., B. J. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2727–2736). Melbourne, Australia: Association for Computational Linguistics.

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

Hammer, A., Goebel, R., Schwarzbach, J., Münte, T. F., & Jansma, B. M. (2007). When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research*, *1146*, 185-198.

Hammer, A., Jansma, B. M., Tempelmann, C., & Münte, T. F. (2011). Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology*, *2*, 1-9.

Hauk, O., Coutout, C., Holden, A., & Chen, Y. (2012). The time-course of single-word reading: Evidence from fast behavioral and brain responses. *NeuroImage*, *60*, 1462–1477.

Heim, S., Opitz, B., & Friederici, A. D. (2002). Broca's area in the human brain is involved in the selection of grammatical gender for language production: Evidence from event-related functional magnetic resonance imaging. *Neuroscience Letters*, *328*, 101–104.

Hobbs, J. (1977). Resolving pronouns. In *Readings in natural language processing*. Los Altos, California, USA.: Morgan Kaufman Publishers, Inc.

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

Indefrey, P., & Levelt, W. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*, 101-144.

Jackson, R. L., Rogers, T. T., & Lambon Ralph, M. A. (2021). Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature Human Behaviour*. (`https://doi.org/10.1038/s41562-020-01034-z`)

Joshi, M., Levy, O., Weld, S., Daniel, & Zettlemoyer, L. (2019). Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5803–5808).

Kietzmann, T., Spoerer, C., Sorensen, L., Cichy, R., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*, 21854–21863.

Kolhatkar, V., & Hirst, G. (2014). Resolving shell nouns. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (p. 499–510). Association for Computational Linguistics.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

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

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

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (p. 188-197). Copenhagen, Denmark: Association for Computational Linguistics.

Lee, K., He, L., & Zettlemoyer, L. (2018). 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)* (p. 687-692). New Orleans, Louisiana: Association for Computational Linguistics.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). 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)* (pp. 138–143). Association for Computational Linguistics.

Longe, O., Randall, B., Stamatakis, E., & Tyler, L. (2007). Grammatical categories in the brain: The role of morphological structure. *Cerebral Cortex*, *17*, 1812–1820.

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

Martin, E., Andrea, & Doumas, A. A., Leonidas. (2019). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B*, *375*, 20190306.

Matchin, W., Sprouse, J., & Hickok, G. (2014). A structural distance effect for backward anaphora in Broca's area: An fMRI study. *Brain and Language*, *138*, 1-11.

Matthew, E. P., Neumann, M., Mohit, I., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of naacl-hlt* (pp. 2227–2237).

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

Miceli, G., Turriziani, P., Caltagirone, C., Capasso, R., Tomaiuolo, F., & Caramazza, A. (2002). The neural correlates of grammatical gender: An fMRI investigation. *Journal of Cognitive Neuroscience*, *14*, 618-628.

Nieuwland, M. S., Petersson, K. M., & Van Berkum, J. A. (2007). On sense and reference: Examining the functional neuroanatomy of referential processing. *Neuroimage*, *37*, 993-1004.

Nieuwland, M. S., & Van Berkum, J. A. (2006). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, *1118*, 155-67.

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

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

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

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

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

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing [Computer software manual]. (bioRxiv 2020.06.26.174482: `https://doi.org/10.1101/2020.06.26.174482`)

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). 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* (p. 102-107). Association for Computational Linguistics.

Toneva, M., Mitchell, T. M., & Wehbe, L. (2021). Combining computational controls with natural text reveals new aspects of meaning composition [Computer software manual]. (bioRxiv 2020.09.28.316935: `https://doi.org/10.1101/2020.09.28.316935`)

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*, 14954—14964.

van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158-71.

van Rij, J., van Rijn, H., & Hendriks, P. (2013). How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, *5*, 564-580.

Vaswani, A., Shazeer, N., Parmar, J., Niki andUszkoreit, Jones, L., Gomez, A. N., Kaiser, Ł., &

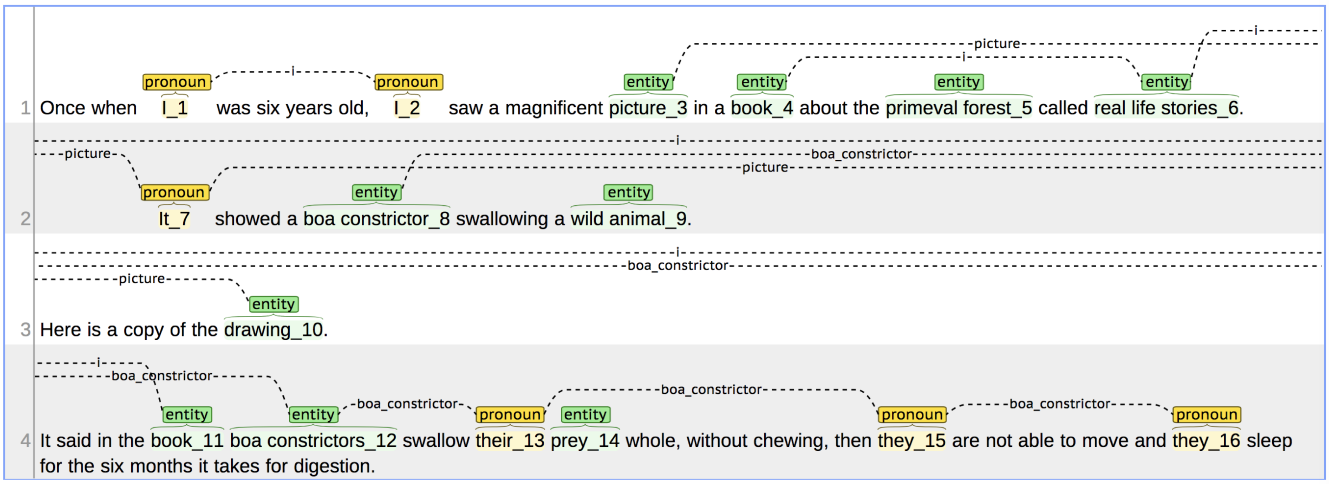
Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems (nips17)* (p. 6000—6010).

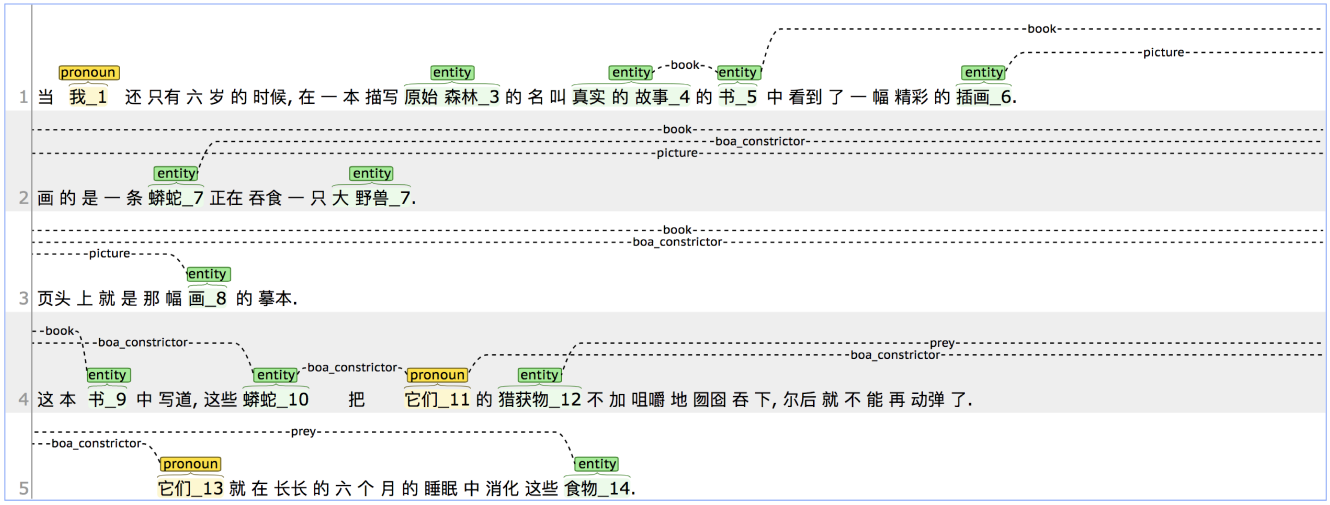
Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Empirical methods in natural language processing* (pp. 233–243).

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

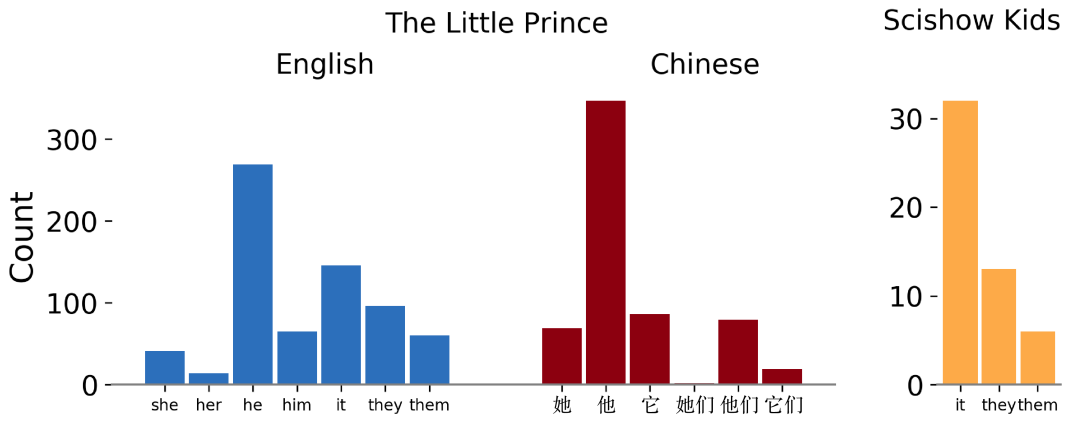
小王子网站 [Computer software manual]. (2020). (`http://www.xiaowangzi.org/`)

Supplementary Information





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 (Stenetorp et al., 2012).



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