The Role of Syntax during Pronoun Resolution: Evidence from fMRI

Jixing Li
Department of Linguistics
Cornell University
jl2939@cornell.edu

Murielle Fabre
Department of Linguistics
Cornell University
mf684@cornell.edu

Wen-Ming Luh
Cornell MRI Facility
Cornell University
wl358@cornell.edu

John Hale
Department of Linguistics
Cornell University
jthale@cornell.edu

Abstract

The current study examined the role of syntactic structure during pronoun resolution. We correlated complexity measures derived by the syntax-sensitive Hobbs algorithm and a neural network model for pronoun resolution with brain activity of participants listening to an audiobook during fMRI recording. Compared to the neural network model, the Hobbs algorithm is associated with larger clusters of brain activation in a network including the left Broca's area.

1 Introduction

Approaching the issue of pronoun resolution from the perspectives of generative linguistics, possible antecedents for pronouns and reflexives are constrained by syntactic structures. For instance, the classical Binding Theory (Chomsky, 1981) states that reflexives are bound in their "local domain" while pronouns are not. [1] For example, "himself" in (1) has to refer to the subject of the inflectional phrase (IP) "Bill", while "him" in (2) cannot refer to "Bill".

(1) John$_i$ thinks that [$_{IP}$Bill$_j$ always criticizes himself$_{*i/j/*k}$].

(2) John$_i$ thinks that [$_{IP}$Bill$_j$ always criticizes him$_{i/*j/k}$].

Nevertheless, it is still unclear what role the binding theory play in the cognitive process of pronoun resolution. It has been argued that explicit syntactic structure and the associated parsing algorithms may not be necessary during sentence comprehension (e.g. Frank and Christiansen, 2018). Furthermore, recent neural network models of coreference resolution (e.g. Clark and Manning, 2016) achieved state-of-the-art results with no explicit syntactic information.

The current study examined the role of syntactic information during pronoun resolution by correlating a complexity measure derived by the syntax-sensitive Hobbs algorithm (Hobbs, 1977) for pronoun resolution with brain activity of participants listened to an audiobook during fMRI recoding. The Hobbs algorithm searches for the gender and number matching antecedent by traversing the parsed syntactic tree in a left-to-right, breadth-first order. We compared brain activation associated with the Hobbs algorithm to that associated with a neural network model for coreference resolution (Clark and Manning, 2016) which encodes no explicit syntactic structures. The results revealed larger clusters for the Hobbs algorithm than for the neural network model in the left Broca's area, the bilateral Angular Gyrus, the left Inferior Temporal Gyrus and the left Precuneus. Given the elements in the Hobbs algorithm including syntactic constraints and gender/number matching, we interpret these areas as supporting morpho-syntactic processing during pronoun resolution.

In the following sections, we briefly describe the Hobbs algorithm and the neural network model and compare their performance on the text of the audiobook. We then describe our linking hypotheses for correlating the models with brain activity, before presenting the methods, results and discussion of the fMRI experiment.

[1] A "local domain" can be roughly defined as the smallest IP or NP which contains the predicate that assigns the theta roles, the complements to which the internal theta roles are assigned, and the subject to which the external theta role is assigned.

Proceedings of the Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing, pages 56–64
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

2 The Hobbs Algorithm

The Hobbs algorithm, originally presented in Hobbs (1977), depends only on a syntactic parser plus a morphological gender and number checker. The input to the Hobbs algorithm includes the target pronoun and the parsed trees for the current and previous sentences. The algorithm searches for a gender and number matching antecedent by traversing the tree in a left-to-right, breadth-first order, giving preference to closer antecedents. 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 for the antecedent in the left of the NP (Principle B: Step 3) and does not go below any NP or S node encountered (Principle A: Step 8). It also respects gender, person, and number agreement, and captures recency and grammatical role preferences in the order it performs the search. Hobbs (1977) evaluated his algorithm on 300 examples containing third person pronouns, and it worked in 88.3% of the cases. With some selectional constraints on dates and location antecedents (i.e., restricting dates and location NPs such as “2018” and “school” to be the antecedent of “it”), the algorithm achieved 91.7% accuracy. However, the test dataset was limited in size and the performance degraded when there were competing antecedents. We propose here to test its accuracy on a larger dataset including 1499 sentences with 465 third person pronouns.

3 The Neural Network Model

The neural network model for pronoun resolution is adapted from the neural network model for both pronominal and nominal coreference resolution (Clark and Manning, 2016). This model consists of a *mention-pair encoder*, a *cluster-pair encoder*, a *mention-ranking model* and a *cluster-ranking model*. The *mention-pair encoder* generates distributed representations for pronoun-antecedent pairs, or mention pairs, by passing relevant features through a feed-forward neural network. The *cluster-pair encoder* generates distributed representations for pairs of clusters through a pooling operation over representations of relevant mention pairs. The *mention-ranking model* scores the candidate antecedents to prune the set of possible antecedent and the *cluster-ranking model* scores coreference compatibility for each pair of clusters.

The input layer of the neural network model consists of a large set of features including word embeddings for the mention pairs, type and length of the mentions, linear distance between the mention pairs, etc. (see Table 1). These feature vectors are concatenated to produce an *I*-dimensional vector $h_0(a, m)$ as the representation for the mention *m* and the antecedent *a*. The input layer then passes through three hidden layers of rectified linear units (ReLU), and the output of the last hidden layer is the vector representation for the mention pair $r_m(a, m)$.

$$h_i(a, m) = ReLU(W_i h_{i-1}(a, m) + b_i)$$

For pairs of clusters $c_i = \{m_1^i, m_2^i, ..., m_{c_i}^i\}$ and $c_j = \{m_1^j, m_2^j, ..., m_{c_j}^j\}$, the *cluster-pair encoder* first forms a matrix $R_m(c_i, c_j) = [r_m(m_1^i, m_1^j), r_m(m_2^i, m_2^j), ..., r_m(m_{c_i}^i, m_{c_j}^j)]$, then applies a pooling operation over $R_m(c_i, c_j)$ to produce a distributed representation for the cluster pair $r_c(c_i, c_j)$. The *mention-ranking model* assigns a score for each mention pair by applying a single fully connected layer of size one the mention pair representation $r_m(a, m)$. The model is then trained with the max-margin training objective.

$$s_m(a, m) = W_m r_m(a, m) + b_m$$

Similarly, the *cluster-ranking model* assigns a coreference score for each cluster pair and an anaphoricity score for mention *m* (i.e., how likely mention *m* has an antecedent). These scores are used to decide whether mention *m* should be merged with one preceding cluster or not during testing.

$$s_c(c_i, c_j) = W_c r_c(c_i, c_j) + b_c$$

$$s_{NA}(m) = W_{NA} r_m(NA, m) + b_{NA}$$

Feature Type	Description
Word embedding	head word
	dependency parent
	first word
	last word
	two preceding words
	two following words
	averaged of the five preceding words
	averaged of five following words
	all words in the mention
	all words in the mention's sentence
	and all words in the mention's document
Mention	type (pronoun/noun/proper name/list)
	position in the document
	contained in another mention or not
	length of the mention in words
Document	genre (broadcast news/newswire/web data)
Distance	intervening sentences
	number intervening mentions
	mentions overlap or not
String matching	head match
	exact string match
	partial string match

Table 1: *Feature set of the neural network model (Clark and Manning, 2016).*

The neural network model encodes no explicit syntactic structures, but it captures semantic information in its word embedding features. It also incorporates discourse-level information such as linear distance between the mention pairs across several sentences, discourse genre, etc. Clark and Manning (2016) trained the model on the CoNLL-2012 Shared Task (Pradhan et al., 2012) and it achieved state-of-the-art results in both the English and Chinese test set.

The neural network model was evaluated on both pronominal and nominal coreference resolution, however, pronouns and full noun phrases (NPs) may rely on very different set of features. For example, string matching and measures for semantic similarity are powerful features for nominal coreference resolution, but are not applicable for pronoun resolution as word embeddings do not represent pronouns well. In addition, it has been argued that pronouns serve a different discourse function from that of full NPs in that full NPs introduce new entities in the discourse and pronouns maintain the reference (Sanford et al., 1988). Based on these arguments, it is reasonable to say that pronoun resolution and full NP coreference involves different cognitive processes.

4 Evaluating the Models on Text Data

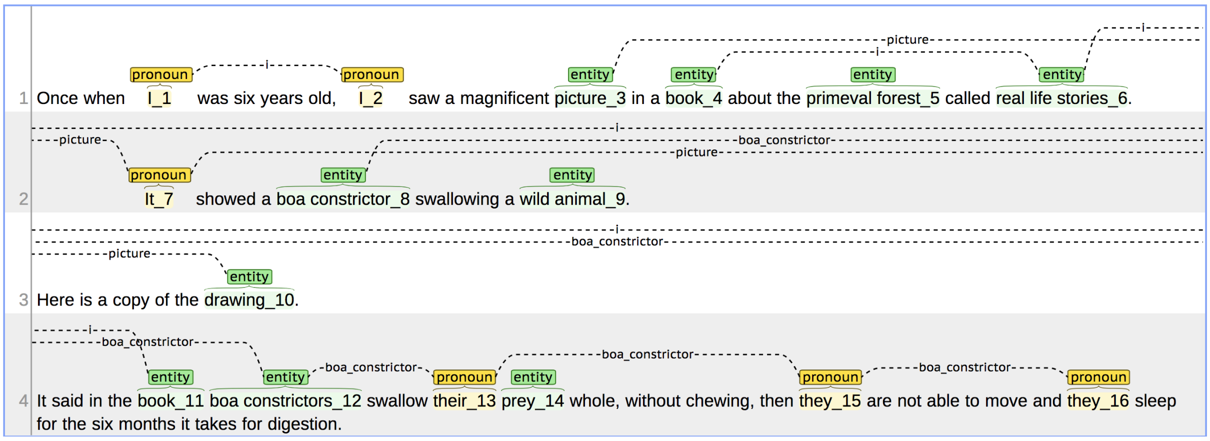
4.1 Text Data

The text data is an English audiobook version of Antoine de Saint-Exupéry's *The Little Prince*. Within the audiobook text, 1755 pronouns and 3127 non-pronominal entities (4882 mentions in total) are identified using the annotation tool brat (Stenetorp et al., 2012; see Figure 1). Reflexives (e.g., "herself") and possessives (e.g., "his") are excluded from the dataset as they have different "binding domains" from pronouns according to the Binding Theory and hence influences performance of the Hobbs algorithm. Pronouns with sentential antecedents (e.g, the second "it" in the conversation "That is funny where you live a day only last a minute." "It is not funny at all."), as well as dummy pronouns (e.g., "it" in "It said in the book that ...") are also removed. The resulting dataset contains 645 first person pronouns, 302 second person pronouns and 675 third person pronouns (see Table 2).

1st	**i**	**me**	**we**	**us**
	505	121	16	3
2nd	**you**			
	302			
3rd	**she**	**her**	**he**	**him**
	41	14	268	64
	it	**they**	**them**	
	136	94	58	

Table 2: Attestations of each pronoun type in *The Little Prince*.

We decided to focus only on the third person pronouns because they provide gender and number information that feeds the Hobbs algorithm. In addition, third person pronouns have been suggested to differ from first and second person pronouns in that first and second person pronouns mark proximity in space and third person pronouns are further away (Ariel, 1990). Therefore, we further excluded third person pronouns whose antecedents



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.

Figure 1: Sample annotations of pronouns and non-pronoun mentions in English, visualized using the annotation tool brat (Stenetorp et al., 2012).

are first and second person pronouns. The final test set contains 465 third person pronouns.

4.2 Model Performance

To evaluate performance of the Hobbs algorithm and the neural network model for third person pronoun resolution in *The Little Prince*, we compared the predicted antecedents for the 465 third person pronouns with the correct immediate antecedents. We considers only the immediate antecedent as the Hobbs algorithm only propose one antecedent and does not group the proposed antecedent into clusters. The syntactic trees for the sentences in the text are parsed by the Stanford PCFG parser (Klein and Manning, 2003).

For the neural network model, we used the pre-trained weights from Clark and Manning (2016) to output a coreference score for all the potential pronoun-antecedent pairs. If the score of the immediate antecedent ranks among top three of all the candidate antecedents, the prediction is marked as correct.

Table 3 shows the accuracy of the Hobbs algorithm and the neural network model for third person pronouns in *The Little Prince*. The neural coreference model only achieves a 0.4 accuracy. Compared with the high F1 score (0.74) for pronoun and full NP coreference resolution on the CoNLL-2012 English test data (Clark and Manning, 2016), this low accuracy confirmed that pronominal and nominal coreference resolution rely on different feature sets. String matching and semantic similarity, for example, may be less powerful for pronominal resolution.

On the other hand, the Hobbs algorithm identifies the correct immediate antecedent for 60% of the third person pronouns. Given the elements of the Hobbs algorithm, it is suggested that linguistically motivated features, especially syntactic constraints and gender/number cues, may be more relevant for third person pronoun resolution in English.

	Accuracy
Hobbs Algorithm	0.60
Neural Network	0.40

Table 3: Performance of the Hobbs algorithm and the neural network model on third pronoun resolution in *The Little Prince*.

4.3 Error Analysis

To probe why the neural network model performed relatively poor than the Hobbs algorithm for third person pronoun resolution, we further divided the dataset into “same sentence” and “different sentence” conditions depending on whether the antecedent occurs within the same sentence of the pronoun. 155 of the 465 third person pronouns have antecedents in the same sentence. Table 4 lists the accuracy of the two models in the two conditions. It can be seen that the Hobbs algorithm performs equally well for the same and different sentence conditions, whereas the neural network model performs worse if the antecedent is not in the same sentence as the pronoun.

A closer examination on the wrong 279 cases predicted by the neural network model revealed that the model tends to be misled by the “partial string match” feature, such that it gives high coref-

	Hobbs	**Neural Network**
Same Sentence	0.60	0.50
Different Sentence	0.60	0.35

Table 4: Accuracy of the Hobbs algorithm and the neural network model for third person pronouns that have antecedents in the same or different sentences.

erence scores for "that" and "they". This confirmed our hypothesis that pronominal and nominal coreference resolution rely on different set of features.

5 Correlating Model Prediction with Brain Activity

5.1 Linking Hypotheses

To explain how the model performance are specifically brought to bear on brain activity, we further correlated activation levels of the antecedents with fMRI time-courses when participants listened to *The Little Prince* in the scanner.

We first selected the 277 third person pronouns whose antecedents are correctly predicted by the Hobbs algorithm, i.e., the true positives, and we calculated the Hobbs distance for each of the 277 pronouns, namely, the number of NPs that the Hobbs algorithm skips before the antecedent NP is proposed. Our linking hypotheses is that a higher Hobbs distance induces a processing effort for pronoun resolution, hence higher hemodynamic response.

Note that the Hobbs distance is different from the number of NP nodes between the pronoun and the antecedents, as the Hobbs algorithm always searches the antecedent to the left of the pronoun in a left-to-right, breadth-first order. Figure 2 shows the Hobbs distance for the two "they" in the example sentence. The immediate antecedent for "they_1" is "their", and the Hobbs distance between "their" and "they_1" is 2 because the algorithm skips the NP "boa constrictors" before proposes "their" as the antecedent. The Hobbs distance for "they_2" is 1 because the correct antecedent is the first proposal by the algorithm.

In comparison, we recorded the coreference score $S_m(a, m)$ generated by the neural network model for the 277 pronouns that correctly predicted by the Hobbs algorithm. We took the negative of the score as a complexity measure for the neural coreference model: the higher the score, the more difficult to retrieve the antecedent. Pearson's *r* revealed no significant correlation between the Hobbs distance and the negative neural coreference score for the 227 third person pronouns ($r = 0.05, p = 0.43$).

5.2 Predicted Brain Activation

Based on the elements in the Hobbs algorithm and the neural network model, we expected the difficulty of pronoun resolution modeled by the Hobbs distance and the neural coreference score to tease apart brain areas that are associated with syntactic and morphological processing, and brain areas that are sensitive to semantic and discourse-level information.

Previous neuroimaging results on pronoun resolution have reported the bilateral Inferior Frontal Gyrus (IFG), the left Medial Frontal Gyrus (MFG) and the bilateral Supramarginal/Angular Gyrus in gender mismatch between pronoun and antecedent (Hammer et al., 2007). We therefore expect activity in these regions for the Hobbs distance metric. We also expect to see activity in the bilateral Superior Temporal Gyrus (STGs) as they have been associated with long distance pronoun-antecedent linking (Matchin et al., 2014). These regions could be relevant for both the Hobbs distance and neural coreference score as they both incorporate some form of "distance" between the pronoun-antecedent pairs. The Precuneus cortex may also be activated with pronouns in general as it has been suggested to track different sorts of story characters (Wehbe et al., 2014).

6 Brain Data

6.1 Participants

Participants were 49 healthy, right-handed, young adults (30 female, mean age = 21.3, range = 18-37). They self-identified as native English speakers, and had no history of psychiatric, neurological or other medical illness that could compromise cognitive functions. All participants were paid for, and gave written informed consent prior to participation, in accordance with the guidelines of the Human Research Participant Protection Program at Cornell University.

6.2 Stimuli

The stimulus was an audiobook version of Antoine de Saint-Exupéry's *The Little Prince*, translated

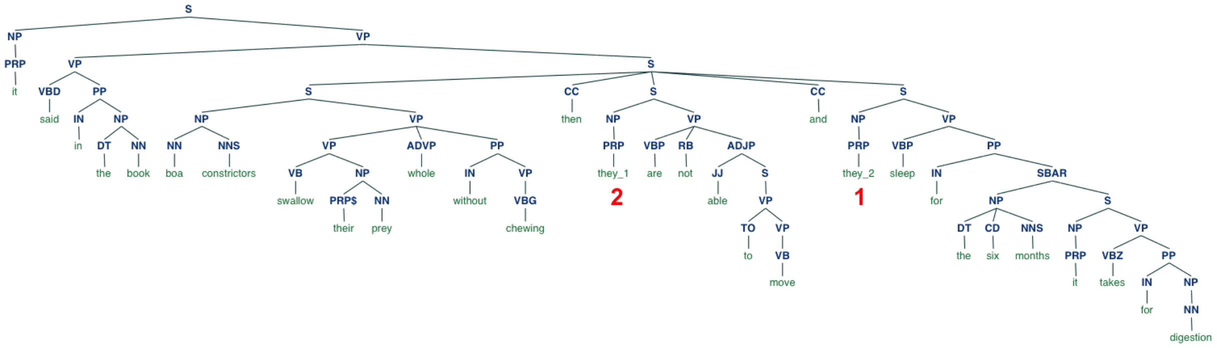


Figure 2: Demonstration of Hobbs distance for third person pronouns in a sentence. The red numbers below the pronouns indicates the Hobbs distance.

by David Wilkinson and read by Nadine Eckert-Boulet. This text contains 3127 non-pronominal mentions and 645 first person pronouns, 302 second person pronouns and 675 third person pronouns (see Table 2). Following the pruning criteria described in Section 5, the final set of data include 277 third person pronouns.

6.3 Procedure

After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner. The presentation script was written in PsychoPy (Peirce, 2007). Auditory stimuli were delivered through MRI-safe, high-fidelity headphones (Confon HP-VS01, MR Confon, Magdeburg, Germany) 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 audiobook lasts for about 94 minutes, and was divided into nine sections, each lasts 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 lasted around 2.5 hours.

6.4 MRI Data Collection and Preprocessing

The 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; TE's=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 is 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.

6.5 Statistical Analysis

At the single subject level, the observed BOLD time course in each voxel were modeled by the difficulty of pronoun resolution derived by the Hobbs Algorithm and the Neural Network Model for third person pronouns time-locked at the offset of each third person pronoun in the audiobook. To further examine the status of Hobbs and Neural Network Models as cognitive models for pronoun resolution, we also included a binary regressor that simply marks the presence of a third person pronoun time-locked at the offset of each third person pronoun in the audiobook.

In addition, three control variables of non-theoretical interest were included in the GLM analysis: *RMS intensity* at every 10 ms of the audio; *word rate* at the offset of each spoken word in time; *frequency* of the individual words in Google Book unigrams [2]. These regressors were added to ensure that any conclusions about pronoun resolution would be specific to those processes, as opposed to more general aspects of speech perception.

At the group level, the activation maps for the Hobbs, neural network and binary regressor were computed using one sample t-test. The voxelwise threshold was set at $p \leq 0.05\ FWE$, with an extent threshold of 50 contiguous voxels ($k \geq 50$).

7 fMRI Results

The largest clusters for the binary third person pronoun regressor were observed in the bilateral Superior Temporal Gyrus (STGs), the left Inferior Frontal Gyrus (IFG), the left Superior Frontal Gyrus (STG), the right Cerebellum and the right Angular Gyrus ($p < 0.05\ FWE$; see Figure 3a).

Hobbs algorithm shows significant activation in the left Precuneus, the bilateral Angular Gyrus, the left IFG and the left SFG ($p < 0.05\ FWE$; see Figure 3b). For the neural network model, although the cluster size is relatively small at the corrected threshold, it has significant clusters in the right STG and the left Middle Temporal Gyrus (MTG; $p < 0.05\ FWE$; see Figure 3c). Table 5 lists all the significant clusters using region names from the Harvard-Oxford Cortical Structure Atlas.

8 Discussion

Activation map for third person pronoun resolution modeled by the Hobbs distance is a subset of the activation map for the binary third person pronoun regressor. Additional activity is observed in the Precuneus for the Hobbs regressor, suggesting that the Precuneus is involved in the process of pronoun-antecedent linking, consistent with Wehbe et al.'s (2014) finding that the Precuneus tracks the characters in a story.

Only the Hobbs algorithm showed an increased activation in the left Broca's area, which has been recurrently reported as correlating with syntactic processing cost linked to antecedent pronoun (Santi and Grodzinsky, 2012), and particularly to the distance between the antecedent and the pronoun (Matchin et al., 2014; Santi and Grodzinsky, 2007).

The bilateral Angular Gyrus activity was also significant for the Hobbs algorithm. Notably, previous literature on German gender agreement in anaphoric reference reported increased activation in the left Angular Gyrus (BA 39) for incongruent biological gender matching (Hammer et al., 2007). Our results supported the role of morphosyntactic processing for gender matching during pronoun resolution at the Angular Gyrus.

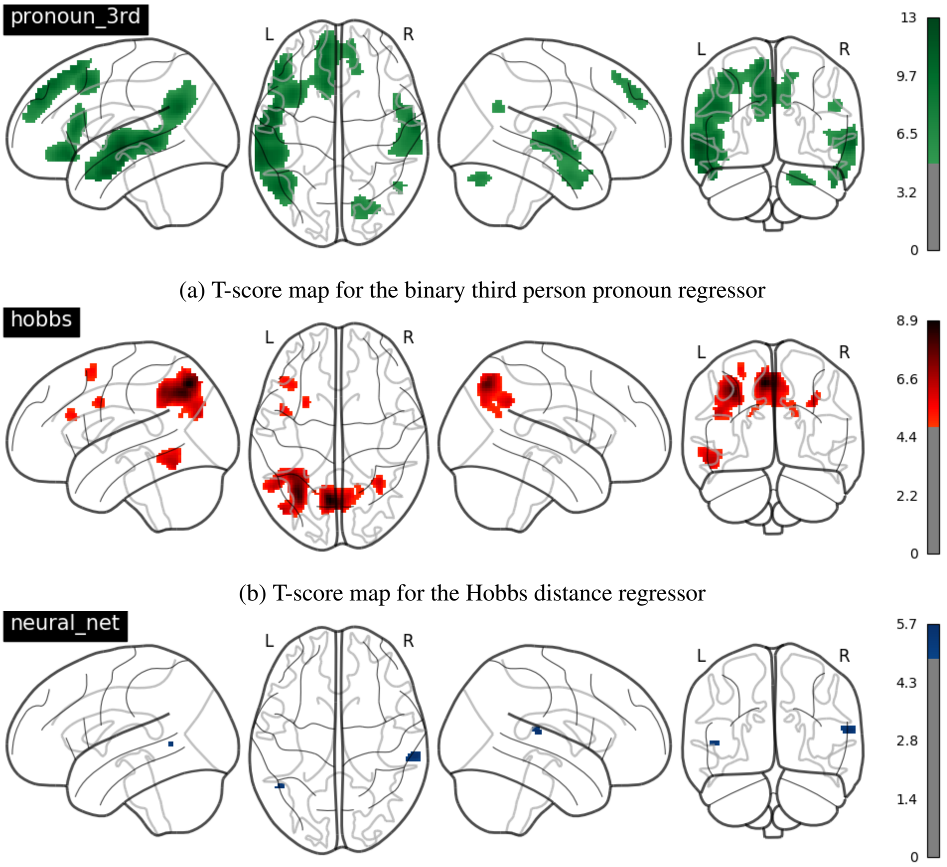
The neural network model encodes different brain activity patterns at the right STG and the left MTG, although the cluster size is relatively small at the corrected threshold. The right STG has been reported to encode linear distance between pronouns and antecedents (Hammer et al., 2007, 2011) and for long distance back anaphora compared to short-distance back anaphora (Matchin et al., 2014). The MTGs have been associated with intra-sentential co-referential link (Fabre, 2017). This is expected as the neural network model encodes the linear distance between the pronoun and the antecedent. The MTGs were also reported to respond to highly predictive lexical access (Fruchter et al., 2015), suggesting that difficulty of pronoun resolution modeled by the neural network scores is likely to involve lexical semantic processing.

9 Conclusion

Comparison of model performance between the Hobbs algorithm and the neural network model on pronoun resolution suggest an important role for syntactic and morphological cues during pronoun resolution. These two types of information were integrated in the Hobbs distance measure that reflects processing difficulty of pronoun resolution. This difficulty measure is associated with significant activity in the left Broca's area, the bilateral Angular Gyrus and the left IFG — a network that has been reported in the neuroimaging literature for anaphora resolution.

Overall, our results show that crossing computational approach and naturalistic stimuli is a promising perspective in neuroimaging to tease apart strongly interwoven cognitive processes. As such, they pave the way for increasing cross-fertilization between computational linguistics and the cognitive neuroscience of language.

[2] http://books.google.com/ngrams



(a) T-score map for the binary third person pronoun regressor

(b) T-score map for the Hobbs distance regressor

(c) T-score map for the negative neural network score regressor

Figure 3: Whole-brain effect with significant clusters for (a) binary third person pronouns effect, (b) difficulty for third pronoun resolution based on the Hobbs algorithm and (c) difficulty for third person pronoun resolution based on the neural coreference model. All images underwent *FWE* voxel correction for multiple comparisons with p $<$ 0.05.

	MNI coordinates			Region	p-value	k-size	t-score
	x	y	z		*FWE-corr*	*cluster*	*peak*
Third Person Pronoun (binary)	-60	-12	-6	left Superior Temporal Gyrus	$<$ 0.001	4411	12.92
	64	-10	-2	right Superior Temporal Gyrus	$<$ 0.001	1625	10.95
	-46	30	-12	left Inferior Frontal Gyrus	$<$ 0.001	706	10.53
	-10	42	46	left Superior Frontal Gyrus	$<$ 0.001	2394	10.45
	18	-74	-30	right Cerebellum	$<$ 0.001	283	7.15
	52	-60	26	right Angular Gyrus	0.004	68	5.84
Hobbs Algorithm	-6	-68	50	left Precuneus	$<$ 0.001	1163	8.86
	-32	-62	42	left Angular Gyrus	$<$ 0.001	1216	8.42
	-52	-56	-16	left Inferior Temporal Gyrus	$<$ 0.001	285	6.54
	34	-52	34	right Angular Gyrus	0.001	119	6.31
	-44	6	34	left Inferior Frontal Gyrus	0.005	55	5.01
	-26	12	60	left Superior Frontal Gyrus	0.007	62	5.63
Neural Network	62	-28	14	right Superior Temporal Gyrus	0.005	48	5.69
	-46	-54	4	left Middle Temporal Gyrus	0.008	13	5.55

Table 5: Significant clusters of BOLD activation for (a) third person pronouns, (b) difficulty for third person pronoun resolution based on the Hobbs algorithm and (c) difficulty for third person pronoun resolution based on the neural coreference model. Peak activations are given in MNI Coordinates ($p <$ 0.05, *FWE*).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No.1607441.

References

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

Noam Chomsky. 1981. *Lectures on government and binding*. Foris, Dordrecht, Holland.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. *arXiv:1606.01323*.

R. W. Cox. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, 29(3):162–173.

Murielle Fabre. 2017. *The sentence as cognitive object - The neural underpinnings of syntactic complexity in Chinese and French*. Ph.D. thesis, INALCO Paris.

Stefan L. Frank and Morten H. Christiansen. 2018. Hierarchical and sequential processing of language: A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience* . *Language, Cognition and Neuroscience*, pages 1–6.

Joseph Fruchter, Tal Linzen, Masha Westerlund, and Alec Marantz. 2015. Lexical Preactivation in Basic Linguistic Phrases. *Journal of Cognitive Neuroscience*, 27(10):1912–1935.

Anke Hammer, Rainer Goebel, Jens Schwarzbach, Thomas F. Münte, and Bernadette M. Jansma. 2007. When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research*, 1146:185–198.

Anke Hammer, Bernadette M. Jansma, Claus Tempelmann, and Thomas F. Münte. 2011. Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology*, 2.

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

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the association for computational linguistics.*, pages 423–430.

Prantik Kundu, Souheil J. Inati, Jennifer W. Evans, Wen-Ming Luh, and Peter A. Bandettini. 2012. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage*, 60(3):1759–1770.

William Matchin, Jon Sprouse, and Gregory Hickok. 2014. A structural distance effect for backward anaphora in Broca's area: An fMRI study. *Brain and Language*, 138:1–11.

Jonathan W. Peirce. 2007. PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8–13.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Anthony J. Sanford, K. Moar, and Simon C. Garrod. 1988. Proper names as controllers of discourse focus. *Language and speech*, 31(1):43–56.

Andrea Santi and Yosef Grodzinsky. 2007. Taxing working memory with syntax: Bihemispheric modulations. *Human Brain Mapping*, 28(11):1089–1097.

Andrea Santi and Yosef Grodzinsky. 2012. Broca's area and sentence comprehension: A relationship parasitic on dependency, displacement or predictability? *Neuropsychologia*, 50(5):821–832.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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*, pages 102–107. Association for Computational Linguistics.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom M. Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *EMNLP*, pages 233–243.