Grammatical Relations in the Listening Brain

Brian Murphy
Knowledge and Data Engineering
Queen's University Belfast
brian.murphy@qub.ac.uk

Jonathan Brennan
Department of Linguistics
University of Michigan
jobrenn@umich.edu

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

***Abstract*—Using an audio-book as a stimulus, and time-locked hemodynamic response modelling of individual events, we show that grammatical subjects and grammatical objects can be differentiated based on fMRI images acquired immediately after hearing a word. A searchlight analysis suggests that grammatical relation information is encoded in the superior and middle temporal gyrus. This finding confirms earlier suggestions that these regions perform combinatorial grammatical processing during naturalistic language comprehension.**

I. Introduction

The term "grammatical relation" refers to syntactic roles like SUBJECT and OBJECT. Perceiving these roles is a crucial part of human language comprehension because of the way they contribute to meaning; for instance the OBJECT is frequently but not always the entity that undergoes an action. The main result of this paper is a functional localization of this type of information in the brains of participants who attentively listened to the spoken narration of a literary text.

To accomplish this, we model the hemodynamic response function (HRF) of the BOLD signal, based on the precise timing of individual words in the speech stream. We carry out whole-brain as well as searchlight analyses (Kriegeskorte et al., 2006) in order to isolate cortical areas that encode grammatical relations.

By decoding whether listeners heard a SUBJECT or an OBJECT in a continuous stream of speech, we localize a particular type of processing within the brain's language network. Results from this naturalistic listening approach converge well with earlier findings based on short stimulus sentences (Frankland and Greene, 2015). The methodology itself highlights the applicability of tools from speech and natural language processing to the neurobiology of language.

II. Method

The brain images used for classification were acquired during a session with an audio-book task, in which participants simply listened to the first chapter of Alice in Wonderland as recited by Kristen McQuillan. This chapter does not include word-play such as the famous Jabberwocky poem that appears elsewhere in the story. We ensured comprehension by administering a twelve-question quiz about the story immediately after participants emerged from the scanner. Out of a total of 29 people scanned, 7 were excluded from the analysis: two scored below chance on the quiz; three had excessive head movement; and two more were excluded after an initial qualitative check identified them has having imperfect registration alignment.

A. Materials

This audio stimulus is freely-available from `librivox.org` and lasts about 12.4 minutes. In the tokenization that we used, this text comprises 2157 words. Using TurboParser (Martins et al., 2009, 2013) we focused our attention on words that were identified by the automatic parser as syntactic subjects ($N = 237$) or syntactic objects ($N = 96$). This procedure extracts the head word in long groups, such as the SUBJECT noun phrase (NP) in example 1 below. In this example, taken from the stimulus text, the head word is identified in boldface.

(1) $[_{\text{NP}}$ Alice's first **thought** $]$ was that it might belong to one of the doors of the hall

Bender (2013, page 63) provides concise overview of headedness per se. Table I shows particular words from the stimulus text with their attestation counts.

SUBJECT		OBJECT	
73	she	14	it
31	it	6	me
29	I	6	herself
19	Alice	3	poison
14	you	3	key
8	there	3	bats
6	rabbit	2	them
4	they	2	rabbit
3	me	2	nothing
3	her	2	jar

TABLE I
ATTESTATION COUNTS FOR FREQUENT SUBJECTS AND OBJECTS

B. Data collection

Imaging was performed using 3T MRI scanner (Discovery MR750, GE Health-care, Milwaukee, WI) with a 32-channel head coil at the ANONYMOUS MRI Facility.

Blood Oxygen Level Dependent (BOLD) signals were collected from twenty-nine participants. Thirteen participants were scanned using a T2*-weighted echo planar imaging (EPI) sequence with a repetition time of 2000 ms, echo time of 27 ms, flip angle of 77°, image acceleration of 2X, field of view of 216 x 216 mm, and a matrix size of 72 x 72. Under these parameters we obtained 44 oblique slices with 3 mm isotropic voxels. Sixteen participants were scanned with a three-echo EPI sequence where the field of view was 240 x 240 mm resulting in 33 slices with an in-plane resolution of 3.75 mm^2 and thickness 3.8mm. This multi-echo sequence was used for reasons that are not related to the present study. Analyses of this second group were based exclusively on images from the second EPI echo, where the echo time was 27.5 ms. All other parameters were exactly the same. This selection of the second-echo images renders the two sets of functional images as comparable as possible.

C. Preprocessing

Preprocessing was done with SPM8 (Friston et al., 2007). Data were spatially-realigned based on 6-parameter rigid body transformation using the 2^{nd} degree B-spline method. Functional (EPI) and structural (MP-RAGE) images were co-registered via mutual information and functional images were smoothed with a 3 mm isotropic Gaussian filter. We used the ICBM template provided with SPM8 to put our data into MNI stereotaxic coordinates, resampling to 2mm isotropic voxels. The data were high-pass filtered at 1/128 Hz and we discarded the first 10 functional volumes. Participants whose inferred head movements exceeded 1 mm in any direction were excluded. We obtained gray-matter masks using SPM8's included segmentation tool. For three participants, this tool was ineffective and the vbm8 toolbox was used instead.

D. Whole-brain and Searchlight Classifiers

All analysis was carried out in Python, with fMRI specific functionality from the PyMVPA software package (Hanke et al., 2009).

1) Data Normalization: For each participant in the analysis, the 4-D sequence of fMRI BOLD images (372 time-points of TR 2s, by 79x95x68 voxels, 2mm isometric) was imported with the application of the participant-specific gray-matter map. Linear detrending was applied to each voxel time-course to remove slow drifts, before the continuous data was transformed into an event-related dataset. This involved collapsing a weighted sequence of BOLD images, following the auditory offset of the head word in a subject or object noun phrase.

This weighted average BOLD signal for each trial was derived by laying a standard HRF model (double gamma, Glover (1999), as implemented in PyMVPA) of duration 20s over the sequence of 11 TRs with which it overlapped, with the TR weights reflecting the area under the HRF curve for each time bin. The resulting event-related dataset was of dimensionality 324 trials, by v voxels, where v was determined by the gray-matter mask for that individual (on average around 125 thousand voxels). This data was z-scored to normalize bias and data spread across voxels, the trials were then partitioned into 5 sequential folds for later cross-validation, and the number of subject and object trials was equalized within each fold by consecutive reduplication of trials from the smaller class (OBJECT), for a total of 462 trials (231 in each class).

2) Whole Brain Classification: For whole-brain classification, a conventional $L2$-regularized logistic regression was used (with default regularization setting) to discriminate single instances of subjects and objects in individual participant sessions. Supervised selection of the top 500 ANOVA-ranked features was used within a 5-fold cross-validation (this classifier setting was adopted from an earlier study of lexical semantics with similar numbers of trials ANONYMOUS). Classification accuracy was the mean of the subject-vs-object discrimination performance, across the 5 test partitions.

3) Searchlight Classification: For the searchlight analysis the same $L2$-regularized logistic regression over a spherical searchlight of approximate radius 4 voxels including the central searchlight voxel (setting of r=3 in PyMVPA's sphere_searchlight). That corresponds to a searchlight sphere of 14mm and volume 1500mm^3. Again the searchlight analysis was run separately over each of the 22 participant session datasets, yielding sensitivity maps of subject/object decoding performance at each point in gray matter.

To aggregate the searchlight results over all participants we applied a one-sample t-test of difference from a population mean (compared to chance performance of 50% classification accuracy). We do not expect perfect alignment of gray-matter across all participants, so only a small proportion of voxels ($<$ 5%) had sensitivity values across all 22 participants. Consequently we considered all 140 thousand voxel locations for which at least half of the participant group had gray-matter present, and from an alpha level of 0.05 we derived a conservative threshold by Bonferroni correction of 3.6×10^{-7}.

III. Results

The whole-brain accuracy over the 22 participants was variable, with values very close to chance performance of 50%, and a mean across participants of only 50.9%. Given this non-significant result, we did not examine the corresponding linear model weights as a sensitivity map.

On the contrary an examination of the mean classification accuracy over all searchlight locations did yield above chance performance for 21 of 22 participants, with a grand mean of 51.2%, $p < 2.5 \times 10^{-5}$.

On the cross-participant analysis we examined every voxel location for which more than half of the participant sessions had a searchlight result. As described above, this set of searchlight accuracies (of size $11 \geq s \geq 22$) was tested a Bonferroni corrected alpha level of 0.05. Of the 140 thousand locations examined, only 12 searchlights were significant across the group of participants. Table II gives MNI coordinates for of all these centers, with anatomical characterizations taken from the Harvard-Oxford Cortical Structure Atlas.

As the localization Figure 1 (overleaf) illustrates, statistically significant discrimination of grammatical subjects and objects across the group of participants can be seen in three broad areas.

The first of these is the temporal lobe of the left hemisphere. This area includes two specific centers. One is centered on a voxel in superior temporal gyrus. Another is located in middle temporal gyrus.

The second general area is frontal in nature, on either side of the longitudinal fissure. This area encompasses six specific locations, four of which are in the frontal pole proper, and two in paracingulate cortex.

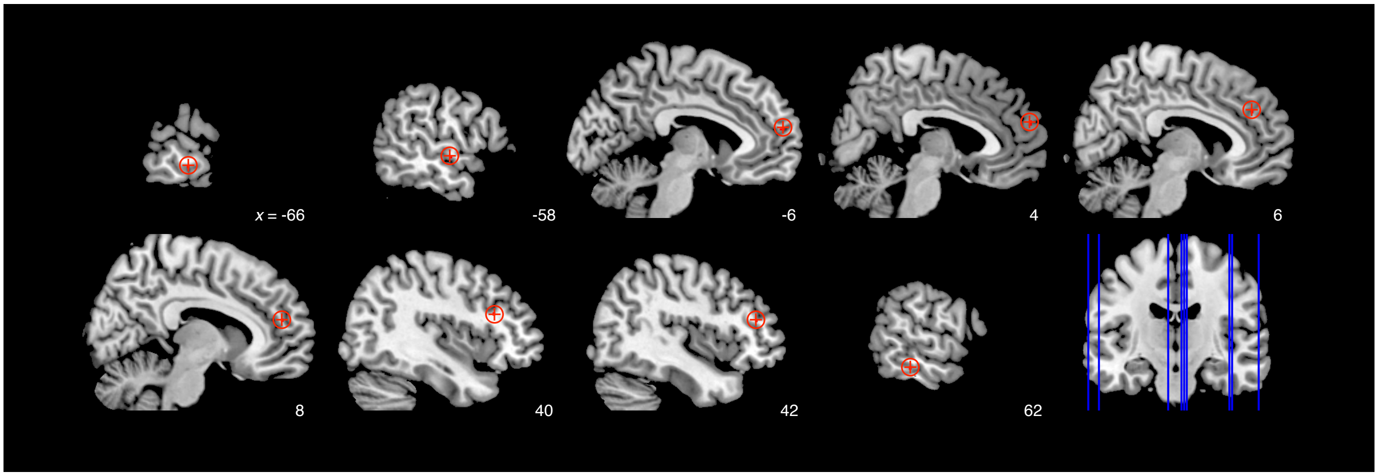

Fig. 1. Location in superior temporal gyrus where classification of subject vs object was above chance.

-58	-10	-4	LH superior temporal gyrus
-66	-22	-10	LH middle temporal gyrus
-6	56,58	16	RH frontal pole, SFG or paracingulate
4	58	20,22	RH frontal pole, SFG or paracingulate
8	48	18	RH paracingulate
6	40	30	RH paracingulate
62	-30	-18	RH inferior temporal gyrus
42	38	18	RH frontal pole
40	24	22	RH MFG or IFG

TABLE II

MNI COORDINATES WHERE SEARCHLIGHT PERFORMANCE EXCEED CHANCE LEVEL. SEE FIGURE 1 FOR VISUALIZATION.

The third general area includes three specific locations within the right hemisphere. Two were frontal, and one was temporal.

IV. DISCUSSION

It is challenging to detect stable patterns in this data because we have a very rapid event-related design ($\sim$ 3 trials per second), and consequent overlapping of individual HRF activations of interest, which we modeled as lasting up to 20s. In a naturalistic paradigm such as this, the collection of word stimuli is heterogeneous so it is not well suited to condition-based deconvolution. Despite this we were still able to pick out significant differences between SUBJECT and OBJECT trials on both searchlight analyses. Whole-brain analyses were inconclusive, though this might be due to a sub-optimal choice of the classifier or its parameter settings.

The first of the three areas in which we observed above-chance classifiability aligns well with the left mid-superior temporal region reported by Frankland and Green in their 2015 experimental study, which used active and passive stimulus sentences. They found a lateral/medial distinction between *semantic* roles. Our results with *syntactic* roles are also in the middle of temporal cortex, but arranged in a superior/inferior relationship.

The second general area in which we observed above-chance classifiability maps on well to the Protagonist's Perspective Interpreter Network proposed by Mason and Just (2006). This makes sense, given that fully 95 out of the 237 words identified as subjects were "she" or "Alice." These are references to the story's main character.

The third general area comprises right-hemisphere homologues of familiar language areas such as inferior frontal gyrus and inferior temporal gyrus. This accords with earlier observations of right hemisphere activation during naturalistic language tasks, even in right-handed participants (Wehbe et al., 2014).

V. CONCLUSION

Because our stimulus was originally written with literary rather than scientific goals in mind, several potential sources of error must be acknowledged. The first is that labeling of a word as a SUBJECT or an OBJECT is imperfect; it was done by a parser trained on newspaper text as opposed to children's literature of the 1860s. Criteria for subjecthood and objecthood remain controversial despite widespread agreement that these grammatical relations must figure, in some way, into the proper characterization of sentence structure (see e.g. Perlmutter (1983) for an influential view).

The second is alignment. While SPM8 finds a best-fitting solution, we ended up discarding datasets whose alignment was clearly off. Classification is done on the basis of participant-specific gray-matter maps which multiply still further the alignment problem.

Finally, the naturalistic character of the stimulus means that potentially co-varying linguistic properties are not well-controlled. For instance, syntactic subjects are frequently (but not always) the semantic AGENT of an action. This type of study cannot determine which, out of many properties that a particular form has, is the true cause of some observed difference.

Even with all these qualifications in mind, our results still invite a simple interpretation: that people perceive words as being subjects or objects during everyday language comprehension, that this ability is subserved by anatomical structures in the temporal lobe, and that these neural populations go into characteristic patterns of activity in response to spoken language. It is these characteristic patterns that allow our classifiers to score above chance.

References

Emily M. Bender. *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*, volume 20 of *Synthesis lectures on human language technologies*. Morgan and Claypool, 2013.

Steven M. Frankland and Joshua D. Greene. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112(37):11732–11737, 2015. doi: 10.1073/pnas.1421236112.

Karl J. Friston, John Ashburner, Stefan J. Kiebel, Thomas E. Nichols, and Wiliam D. Penny, editors. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.

Gary H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4):416–429, 1999. ISSN 1053-8119. doi: 10.1006/nimg.1998.0419.

Michael Hanke, Yaroslav O Halchenko, Per B Sederberg, Emanuele Olivetti, Ingo Fründ, Samuel Garcia, and Université Claude. PyMVPA : a unifying approach to the analysis of neuroscientific data. *Neuroinformatics*, 3(February):1–13, 2009. ISSN 16625196. doi: 10.3389/neuro.11.003.2009.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter A. Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–8, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0600244103.

Andre Martins, Noah Smith, and Eric Xing. Concise integer linear programming formulations for dependency parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 342–350, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Andre Martins, Miguel Almeida, and Noah A. Smith. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Robert A. Mason and Marcel Adam Just. *Neuroimaging contributions to the understanding of discourse processes*, pages 765–799. Elsevier, 2006.

David M. Perlmutter. *Studies in relational grammar*. University of Chicago Press, Chicago, 1983.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.