The Compositional Nature of Event Representations in the Human Brain

Andrei Barbu,[1*] N. Siddharth,[1] Caiming Xiong,[2]
Jason J. Corso,[2] Christiane D. Fellbaum,[3]
Catherine Hanson,[4] Stephen José Hanson,[4]
Sébastien Hélie,[5] Evguenia Malaia,[6] Barak A. Pearlmutter,[7]
Jeffrey Mark Siskind,[1] Thomas Michael Talavage,[1] Ronnie B. Wilbur[8]

[1]School of Electrical and Computer Engineering
Purdue University, West Lafayette IN 47907

[2]Department of Computer Science and Engineering
SUNY Buffalo, Buffalo NY 14260

[3]Department of Computer Science, Princeton University, Princeton NJ 08540

[4]Department of Psychology and Rutgers Brain Imaging Center
Rutgers University, Newark NJ 07102

[5]Psychological Sciences, Purdue University, West Lafayette IN 47907

[6]Southwest Center for Mind, Brain, and Education
University of Texas at Arlington, Arlington TX 76019

[7]Hamilton Institute and Department of Computer Science
National University of Ireland Maynooth, Co. Kildare, Ireland

[8]Department of Speech, Language, and Hearing Sciences and Linguistics Program
Purdue University, West Lafayette IN 47907

[*]To whom correspondence should be addressed; E-mail: andrei@0xab.com.

How does the human brain represent simple compositions of constituents: actors, verbs, objects, directions, and locations? We had subjects view action-sequence videos during neuroimaging (fMRI) sessions and identified lexical descriptions of those videos by decoding the brain representations based only

on their fMRI activation patterns. We independently decoded each constituent from a single stimulus presentation and compared the performance of such independent classification to joint classification of the aggregate concepts. Independent classification of constituents performs largely the same as joint classification as measured by accuracy and correlation. The brain regions used for independent constituent classification are largely disjoint and largely cover those used for joint classification. This allows recovery of sentential descriptions of stimulus videos by composing the results of the independent constituent classifiers.

Introduction

The compositional nature of thought is taken for granted by many in the cognitive-science community. The representations commonly employed compose aggregated concepts from constituent parts (*1–4*). Humans need not employ compositional representations; indeed, many argue that such representations may be ill suited as models of human cognition (*5*). This is because concepts like *verb* or even *object* are human constructs; there is hence debate as to how they arise from percepts (*6*). Recent advances in brain-imaging techniques enable exploration of the compositional nature of brain activity. To that end, subjects underwent functional magnetic resonance imaging (fMRI) while exposed to stimuli that evoke complex brain activity which was decoded, piece by piece. The video stimuli depicted events described by entire sentences composed of an *actor*, a *verb*, an *object*, and a *direction* of motion or a *location* of the event in the field of view. By decoding complex brain activity into its constituent parts, and by further demonstrating that

- accuracy of classification by classifiers trained independently on the constituents is largely the same as that of classifiers trained jointly on constituent pairs and triples,

- the brain regions employed by the per-constituent classifiers are largely pairwise disjoint, and
- the brain regions employed by the joint classifiers largely consist of the unions of the brain regions employed by the component constituent classifiers

we show evidence for the neural basis of the compositionality of event representations. We know of no other work that demonstrates this neural basis by simultaneously decoding the brain activity for all of these constituents from the same video stimulus.

Compositionality can refer to at least two different notions. It can refer to the result of a composition. For example, $2 + 3 = 5$ composes 5 out of 2 and 3. It is impossible to reconstruct 2 and 3 from the result 5. It can also refer to a specification of the structure of the composition. For example, '$2 + 3$'. From such it is possible to extract '2' and '3'. The same issue arises with semantics. *John* combines with *walked* to yield *John walked*. The result of such a composition could be some nondecomposable representation. Yet the structural specification of such a composition could be decomposed into its constituents.

Does the brain employ such decomposable representations? It is conceivable that representations are decomposable at some processing stages but not others. When seeing John walk, neural activity might encode regions in the field of view that reflect the aggregate percept of John walking. Moreover, that aggregate percept might be spread across neural activity in space and/or time. A behavioral response to seeing John walk, such as walking towards him, also might reflect the aggregate percept, not the percepts of John alone or walk alone, because the percept of John sitting or of Mary walking might evoke different responses. Motor response might reflect an aggregate percept which might also be spread across neural activity in space and/or time. Thus there appear to be at least some processing stages, particularly at the inputs and outputs, where representations might not be decomposable. The question is whether there

exist other intermediate processing stages that are. We investigate this question.

Language itself is compositional. Sentences are composed of words. It seems likely that when a percept involves language, either auditory or visual (orthographic, signed), the neural representation of that percept would be decomposable, at least at the input. It also seems likely that when a behavioral response involves language, oral or visual (written, signed), the neural representation of that motor response would be decomposable, at least at the output. It would be surprising, however, if a purely visual task that involved no linguistic stimuli or response would evoke decomposable brain activity. Our experiment design investigates just that. (See the supplementary material for discussion.)

A requirement for decomposable representations is a degree of independence of the constituent representations. It is not possible to recover 2 and 3 from 5 because additional information enters into the process $2 + 3 = 5$, namely addition. It would only be possible to recover '2' and '3' from '$2 + 3$' if their representations were independent. Just as decomposability need not be black and white—there may be both decomposable and nondecomposable representations employed in different brain regions—independence also need not be black and white—degree of independence may vary. We investigate the structural decomposability of aggregate compositional percepts by measuring and demonstrating a high degree of independence of the constituents.

Recent work on decoding brain activity has recovered object class from nouns presented as image, video, aural, and orthographic stimuli (*7–14*). Similar work on verbs has primarily been concerned with identifying active brain regions (*14–18*). Other recent work has demonstrated the ability to decode the actor of an event using personality traits (*19*). These past successes suggest that one can investigate our novel hypothesis using significant extensions of these prior methods combined with several novel analyses.

We investigate compositionality as applied to sentence structure—objects fill argument positions in predicates that combine to form sentential meaning—and decompose such into independent constituents. Recent work has identified brain regions correlated with compositionality that may not be decomposable using a task called *complement coercion* (*20*). Subjects were presented with sentences whose whose meanings were partly implied rather than fully expressed overtly through surface constituents. For example, the sentence *The boy finished the pizza* is understood as meaning that the pizza was eaten, even though the verb *eat* does not appear anywhere in the sentence (*21*). The presence of *pizza*, belonging to the category *food*, coerces the interpretation of *finish* as *finish eating*. By contrast, *He finished the newspaper* induces the interpretation *finish reading*. Because syntactic structure in this prior experiment was held constant, the assumption was that coercion reflects incorporation of additional information in the result that is absent in the constituents. Brain activity measured using magnetoencephalography (MEG) showed activity related to coercion in the anterior midline field. This result suggests that there may be some regions that do not exhibit decomposable brain activity but does not rule out the possibility that there are other regions that do.

Experiment Design

We hypothesized the existence of brain regions that exhibit decomposable brain activity and conducted an experiment to evaluate this hypothesis by demonstrating the ability to decode the brain activity evoked by a complex visual stimulus into a sentence that describes that stimulus by independently decoding the constituent words. Subjects were shown videos that depict sentences of the form: the *actor verb* the *object direction/location*. Subjects were shown sample video prior to imaging and were informed of the structure of the stimuli and the intended collection of actors, verbs, objects, directions, and locations. They were asked to think about the sentence depicted by each video and otherwise had no task.

Videos depicting one of four human actors performing one of three verbs (*carry*, *fold*, and *leave*), each with one of three objects (*chair*, *shirt*, and *tortilla*), on either side of the field of view were filmed for this task. The verbs were chosen to be discriminable based on the following features (*16*):

carry	−state-change	+contact
fold	+state-change	+contact
leave	−state-change	−contact

Nouns were chosen based on categories previously found to be discriminable: *chair* (furniture), *shirt* (clothing), and *tortilla* (food) and also selected to allow each verb to be performed with each noun (*11*). All stimuli enactments were filmed against the same nonvarying background, which contained no other objects except for a table (Fig. 1).

We collected multiple videos, between 3 and 7, for each element of the cross product of actor, verb, and object. Variation in direction of motion and location in the field of view was accomplished by mirroring the videos about the vertical axis. Such mirroring induces variation in direction of motion (leftward *vs.* rightward) for the verbs *carry* and *leave* and induces variation in the location in the field of view where the verb *fold* occurs (left half *vs.* right half).

We employed a rapid event-related design (*11*). We presented 2s video clips at 10fps followed by an average of 4s (minimum 2s) fixation. Each run comprised 72 stimulus presentations spanning 244 captured brain volumes, with eight runs per subject, and ended with 24s of fixation. Each run was individually counterbalanced for each of the four conditions (actor, verb, object, and mirroring). Runs were separated by several minutes, during which no stimuli were presented, no data was gathered, and subjects engaged in unrelated conversation with the experimenters.

Imaging was performed at Purdue University using a 3T GE Signa HDx scanner (Waukesha, Wisconsin) with a Nova Medical (Wilmington, Massachusetts) 16 channel brain array to collect whole-brain volumes via a gradient-echo EPI sequence with 2000ms TR, 22ms TE,

200mm×200mm FOV, and 77° flip angle. We acquired 35 axial slices with a 3.0mm slice thickness using a 64×64 acquisition matrix resulting in 3.125mm×3.125mm×3.0mm voxels.

Analysis

This experiment design supports the following classification analyses for constituents:

actor one-out-of-4 actor identity

verb one-out-of-3 verb (*carry*, *fold*, and *leave*)

object one-out-of-3 noun (*chair*, *shirt*, and *tortilla*)

direction one-out-of-2 direction of motion for *carry* and *leave* (leftward *vs.* rightward)

location one-out-of-2 location in the field of view for *fold* (right *vs.* left)

All analyses reported follow the same procedure and employ the same methods and classifiers. Videos were shown to subjects who were asked to think about some aspect(s) of the video while whole-brain fMRI scans were acquired every two seconds. Because fMRI acquisition times are slow, equal to the length of the video stimuli, a single brain volume that corresponds to the peak brain activation induced by that video stimulus was classified to recover the features that the subjects were asked to think about.

Brain scans were processed using AFNI (*22*) to skull-strip each volume, motion correct and detrend each run, and align all scans for a given subject to a subject-specific reference volume. Voxels within a run were z-scored, subtracting the mean value of that voxel for the run and dividing by its variance. Because each brain volume has very high dimension, 143,360 voxels, we eliminated voxels by computing a per-voxel Fisher score on our training set and keeping the 4,000 highest-scoring voxels. The Fisher score of a voxel v for a classification task with

C classes where each class c has n_c examples is computed as

$$\frac{\sum_{c=1}^{C} n_c(\mu_{c,v} - \mu)^2}{\sum_{c=1}^{C} n_c \sigma_{c,v}^2}$$

where $\mu_{c,v}$ and $\sigma_{c,v}$ are the per-class per-voxel means and variances and μ is the mean for the entire brain volume. The resulting voxels were then analyzed with *Linear Discriminant Dimensionality Reduction* (*23*) to select a smaller number of potentially-relevant voxels, selecting on average 1084 voxels per-subject per-fold. Both stages of voxel selection were performed independently for each fold of each subject only on the training set but applied to the test set prior to classification.

A linear support vector machine (SVM) was employed to classify the selected voxels (*24*). One run was taken as the test set and the remaining runs were taken as the training set. The third brain volume after the onset of each stimulus was taken along with the class of the stimulus to train an SVM. This lag of three brain volumes is required because fMRI does not measure neural activation but instead measures the flow of oxygenated blood, the blood-oxygen-level-dependent (BOLD) signal, which correlates with increased neural activation. It takes roughly five to six seconds for this signal to peak which puts the peak in the third volume after the stimulus presentation. Cross validation was performed by choosing each of the different runs as the test set. The separation between runs allowed runs to constitute folds for cross validation without introducing spurious correlation in brain activity between runs.

To control for the potential of classifying noise or irrelevant features, we determined the brain regions relevant to each analysis using two distinct methods. We first employed a spatial searchlight (*25*) which slides a small sphere across the entire brain volume and repeats the above analysis keeping only the voxels inside that sphere. We used a sphere of radius three voxels, densely placed at the center of every voxel, and did not perform any dimensionality reduction on

the remaining voxels. We then performed an eight-fold cross validation as described above for each position of the sphere. We also back-projected the SVM coefficients onto the anatomical scans—the higher the absolute value of the coefficient the more that voxel contributes to the classification performance of the SVM—and used a classifier (*10*) with a different metric, $w(i)^2$.

Results

We collected data for eight subjects but discarded the data for one subject due to excessive motion. One subject did eight runs without exiting the scanner. All other subjects exited the scanner at various points during the set of eight runs, which required cross-session registration. All subjects were aware of the experiment design, were informed of the intended depiction of each stimulus prior to the scan, and were instructed to think of the intended depiction after each presentation. All analyses below employ identical eight-fold cross-validation across the eight runs per subject.

Fig. 2 presents the per-constituent classification accuracies, both per-subject (top) and aggregated across subject (bottom). We achieve performance well above chance on all five constituents with only a single fold for subject 1 and two folds for subject 2 at chance for the actor analysis and two folds for subject 2 at chance for the location analysis. Average performance across subject is also well above chance (**actor** $33.33\%^{***}$, chance 25.00%; **verb** $78.92\%^{***}$, chance 33.33%; **object** $59.80\%^{***}$, chance 33.33%; **direction** $84.60\%^{***}$, chance 50.00%; **location** $71.28\%^{***}$, chance 50.00%; see the supplementary material for a discussion of statistical significance).

We conducted an additional analysis to measure the independence of the representations for these constituents. We trained classifiers jointly for all constituent pairs, except for verb and location because location only applied to a single verb *fold*, and compared the classification accuracy against independent application of the classifiers trained on the constituents in isolation

(Fig. 3a, top). We similarly trained classifiers jointly for all constituent triples, except for actor, object, and location due to lack of sufficient training data, and performed a similar comparison (Fig. 3a, bottom). An independent classification was deemed correct if it correctly classified all of the constituents in the pair or triple.

We conducted a further analysis to measure the accuracy of decoding an entire sentence from a single stimulus. Training a joint classifier on entire sentences would require a sufficiently large number of samples for each of the 72 possible sentences ($4\times3\times3\times2$), which would be infeasible to gather due to subject fatigue. However, we independently classified each sample with the per-constituent classifiers and combined the results as described above (Fig. 3b). Average performance across subject is well above chance ($13.84\%^{***}$, chance 1.39%).

We quantified the degree of independence of the classifiers by comparing the individual classification results of the independent classifiers to those produced by the joint classifiers, for all constituent pairs and triples, by computing the accuracy and Matthews correlation coefficient (MCC), for multi-class classification (*26*), over the samples where the joint classifier was correct, yielding an average accuracy of 0.7056 and an average correlation of 0.4139 across all analyses (Fig. 4).

To locate brain regions used in the previous analyses, we employed a spatial-searchlight linear-SVM method on all subjects. We used the accuracy to determine the sensitivity of each voxel and thresholded upward to less than 10% of the cross-validation measures. These measures are overlaid and (2-stage) registered to MNI152 2mm anatomicals. We performed this searchlight analysis independently for all of the constituent and joint classifiers. The resulting constituent regions (omitting actor) are color coded according to the specific constituent being decoded. We also back-projected the thresholded SVM coefficients for all constituents, including actor, produced by the analysis in Fig. 2, for all subjects, onto the anatomical scan, aggregated across run. The resulting regions produced by both analyses for subject 1 are shown

in Fig. 5. (Figures for all subjects are included in the supplementary material.)

To further quantify the degree of spatial independence, we compared the brain regions indicated by searchlight and by the thresholded SVM coefficients of the independent classifiers to those of the joint classifiers, for all constituent pairs and triples. We first computed the percentage of voxels in the union of the constituents for the independent classifier that were also in the intersection (Fig. 6 top). We also computed the percentage of voxels in the joint classifier that are shared with the independent classifier (Fig. 6 bottom).

Discussion

Our results indicate that brain activity corresponding to each of the constituents, actor, verb, object, direction, and location, can be reliably decoded from fMRI scans, both individually, and in combination. Given neural activation, we can decode what the subjects are thinking about. We know of no other work that simultaneously decodes the brain activity corresponding to all of these constituents from a single video stimulus.

Furthermore, our analysis indicates that a decomposable neural representation for each of these five constituents exists in the brain. This is surprising; intermediate neural representation could have been all interdependent, just like the inputs and outputs. People engage in distinct motion when *fold*ing *chair*s, *shirt*s, and *tortilla*s. If the representation of a verb, like *fold*, was neurally encoded for a particular object, for example, to reflect the particular motion involved when performing the action denoted by that verb, it would not be possible to decode this verb with performance above chance in our experiment design, because it is counterbalanced with respect to the objects with which the action is being performed. Moreover, if there were some level of object specificity in verbs, one would expect this to be reflected in marked decrease in classification accuracy of independent classifiers for verb and object over a joint classifier for the pair. This, however, does not appear to be the case: averaged across subject, the joint verb-

object classifier has $49.58\%^{***}$ accuracy while the independent one has $47.94\%^{***}$. The relative performance of joint *vs.* independent classification appears similar across all combinations of constituents, not just verbs and objects (Fig. 3), so much so that we can decode an ***entire sentence*** from a single stimulus, with accuracy far above chance, using per-constituent classifiers trained independently on those constituents. Moreover, joint and independent classification are highly correlated (Fig. 4), indicating that the joint classifiers are not making significant use of information beyond that available to the independent classifiers.

In general, our searchlight analysis and our back-projected SVM coefficients (Fig. 5) indicate that such decoding relies on different brain regions for different constituents. **Actor** activity is present in the fusiform face area (*27*). **Verb** activity is present in visual-pathway areas (lateral occipital-LO, lingual gyrus-LG, and fusiform gyrus) as well as prefrontal areas (inferior frontal gyrus, middle frontal gyrus, and cingulate) and areas consistent with the 'mirror system' (*28*) and the so-called 'theory of mind' (pre-central gyrus, angular gyrus-AG, and superior parietal lobule-SPL) areas (*29, 30*). **Object** activity is present in the temporal cortex, and agrees with previous work on object-category encoding (*31*). **Direction** and **location** activity is present in the visual cortex with significant **location** activity occurring in the early visual cortex. More specifically, quantitative analysis of the brain regions indicated by both searchlight and the thresholded SVM coefficients indicates that the brain regions used for independent constituent classification are largely disjoint (3.72% for searchlight and 2.08% for thresholded SVM weights, averaged across both subject and analysis) and largely cover (60.83% for searchlight and 51.16% for thresholded SVM weights, averaged across both subject and analysis) those used for joint classification (Fig. 6).

Compositionality is a rich notion. Not only must it be possible to determine 2 and 3 from '$2 + 3$', it must be possible to determine that 2 is an argument of this addition but not the addition in '$4 + (3 \times 2)$', even though it appears elsewhere in the formula. For language and

vision, it must be possible to determine that a person is folding the chair and not the shirt, when a shirt is present in the field of view but is not being folded. The present analysis can be construed as computational identification of associative-linguistic representations, a form of syntax-less language learning, without PFC engagement (*32*). Further, not all operations are commutative or symmetric: it must be possible to distinguish '$2 - 3$' from '$3 - 2$.' For language and vision, some predicates are also asymmetric; it must be possible to distinguish between a person approaching a dog from a dog approaching a person. Making such distinctions will require analyzing fine-grained prefrontal cortical activity, likely using a region-of-interest approach (*33*). Finally, the individual constituents may themselves be decomposable. Verbs like *raise* and *lower* may decompose into lower-level constituents indicating causation of upward *vs.* downward motion where the lower-level constituents denoting causation and motion are shared between the two verbs but those denoting direction are not (*1, 34, 35*). For now, our findings are agnostic to these issues.

Conclusion

We have demonstrated that it is possible to decode a subject's brain activity into constituents, which when combined, yield a sentential description of a video stimulus. To do so, we conducted the first study which decodes brain activity associated with actors, verbs, objects, directions, and locations from video stimuli, both independently and jointly. Our results are the first to indicate that the neural representations for these constituents compose together to form the meaning of a sentence, apparently without modifying one another, even when evoked by purely visual, nonlinguistic stimuli. These results are in concord with Jackendoff's Cognitive Constraint and Conceptual Structure Hypothesis (*36*) and indicate that representations which attempt to decompose meaning into constituents may have a neural basis.

References and Notes

1. G. A. Miller, P. N. Johnson-Laird, *Language and Perception* (Harvard University Press, Cambridge, MA, 1976).

2. R. Jackendoff, Semantics and cognition (1983). *E.g.*, (10.10a–j) p. 192.

3. S. Pinker, Learnability and cognition (1989). *E.g.*, (5.46) and (5.47) p. 218.

4. S. M. Kosslyn, Image and brain: The resolution of the imagery debate (1996). Second paragraph, p. 6.

5. R. A. Brooks, *Artificial Intelligence* **47**, 139 (1991).

6. B. C. Smith, *On the Origin of Objects* (MIT Press, Cambridge, MA, 1996).

7. A. Puce, T. Allison, M. Asgari, J. C. Gore, G. McCarthy, *The Journal of Neuroscience* **16**, 5205 (1996).

8. S. J. Hanson, T. Matsuka, J. V. Haxby, *Neuroimage* **23**, 156 (2004).

9. Y. Miyawaki, *et al.*, *Neuron* **60**, 915 (2008).

10. S. J. Hanson, Y. O. Halchenko, *Neural Computation* **20**, 486 (2009).

11. M. A. Just, V. L. Cherkassky, S. Aryal, T. M. Mitchell, *PloS One* **5**, e8622 (2010).

12. A. C. Connolly, *et al.*, *The Journal of Neuroscience* **32**, 2608 (2012).

13. F. Pereira, M. Botvinick, G. Detre, *Artificial Intelligence* **194**, 240 (2012).

14. A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, *Neuron* **76**, 1210 (2012).

15. J. W. Kable, A. Chatterjee, *Journal of Cognitive Neuroscience* **18**, 1498 (2006).

16. D. Kemmerer, J. Gonzalez Castillo, T. Talavage, S. Patterson, C. Wiley, *Brain and Language* **107**, 16 (2008).

17. D. Kemmerer, J. Gonzalez Castillo, *Brain and Language* **112**, 54 (2010).

18. Y. Coello, C. Bidet-Ildei, *Language and Action in Cognitive Neuroscience*, Y. Coello, A. Bartolo, eds. (Psychology Press, New York, NY, 2012), chap. 4, pp. 83–110.

19. D. Hassabis, *et al.*, *Cerebral Cortex* **23** (2013).

20. L. Pylkkänen, J. Brennan, D. K. Bemis, *Language and Cognitive Processes* **26**, 1317 (2011).

21. J. Pustejovsky, *Generative Semantics* (MIT Press, Cambridge, MA, 1995).

22. R. W. Cox, *Computers and Biomedical Research* **29**, 162 (1996).

23. Q. Gu, Z. Li, J. Han, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (2011), pp. 549–564.

24. C. Cortes, V. Vapnik, *Machine Learning* **20**, 273 (1995).

25. N. Kriegeskorte, R. Goebel, P. Bandettini, *Proceedings of the National Academy of Sciences of the United States of America* **103**, 3863 (2006).

26. J. Gorodkin, *Computational Biology and Chemistry* **28**, 367 (2004).

27. N. Kanwisher, J. McDermott, M. M. Chun, *Journal of Neuroscience* **17**, 4302 (1997).

28. M. A. Arbib, *Action to Language via the Mirror Neuron System* (Cambridge University Press, Cambridge, UK, 2006).

29. N. F. Dronkers, D. P. Wilkins, R. D. Van Valin, Jr., B. B. Redfern, J. J. Jaeger, *Cognition* **92**, 145 (2004).

30. U. Turken, N. F. Dronkers, *Frontiers in Systems Neuroscience* **5** (2011).

31. M. S. Gazzaniga, R. B. Ivry, G. R. Mangun, *Cognitive Neuroscience: The Biology of the Mind* (W. W. Norton & Company, New York, NY, 2008), third edn.

32. A. D. Friederici, J. L. Mueller, B. Sehm, P. Ragert, *Journal of Cognitive Neuroscience* **25**, 814 (2013).

33. H. A. Jeon, A. D. Friederici, *Nature Communications* **4**, 2041 (2013).

34. R. Jackendoff, *Semantics and Cognition* (MIT Press, Cambridge, MA, 1983).

35. S. Pinker, *Learnability and Cognition* (MIT Press, Cambridge, MA, 1989).

36. R. Jackendoff, Semantics and cognition (1983). See pp. 16–22 including (1.3) and (1.4), also included in the supplementary material.

37. S. M. Kosslyn, *Image and Brain: The Resolution of the Imagery Debate* (MIT Press, Cambridge, MA, 1996).

Acknowledgments

AB, NS, and JMS were supported, in part, by Army Research Laboratory (ARL) Cooperative Agreement W911NF-10-2-0060. CX and JJC were supported, in part, by ARL Cooperative Agreement W911NF-10-2-0062 and NSF CAREER grant IIS-0845282. CDF was supported, in part, by NSF grant CNS-0855157. CH and SJH were supported, in part, by the McDonnell Foundation. BAP was supported, in part, by Science Foundation Ireland grant 09/IN.1/I2637. The views and conclusions contained in this document are

those of the authors and should not be interpreted as representing the official policies, either express or implied, of the supporting institutions. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein. Dr. Gregory G. Tamer, Jr. provided assistance with imaging and analysis.

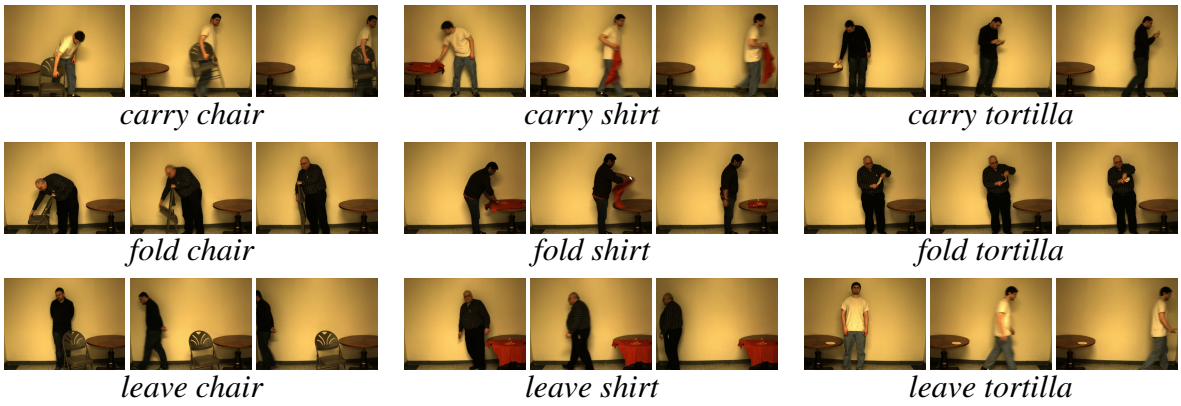


Figure 1: Key frames from sample stimuli. Stimulus videos are included in the supplementary material.

	Classification Accuracy				
Subject	**actor**	**verb**	**object**	**direction**	**location**
1	30.4%	77.6%***	55.4%***	84.6%***	69.8%***
2	31.4%	67.4%***	54.2%***	76.3%***	61.5%*
3	35.6%***	83.0%***	62.3%***	93.0%***	67.7%***
4	35.2%***	81.6%***	66.0%***	82.3%***	73.4%***
5	33.2%**	87.0%***	66.3%***	88.0%***	75.5%***
6	32.3%**	77.6%***	57.5%***	80.7%***	79.7%***
7	35.2%***	78.3%***	56.9%***	87.2%***	71.4%***
mean	33.33%***	78.92%***	59.80%***	84.60%***	71.28%***
stddev	4.46	7.88	6.66	7.57	10.97
chance	25.00%	33.33%	33.33%	50.00%	50.00%

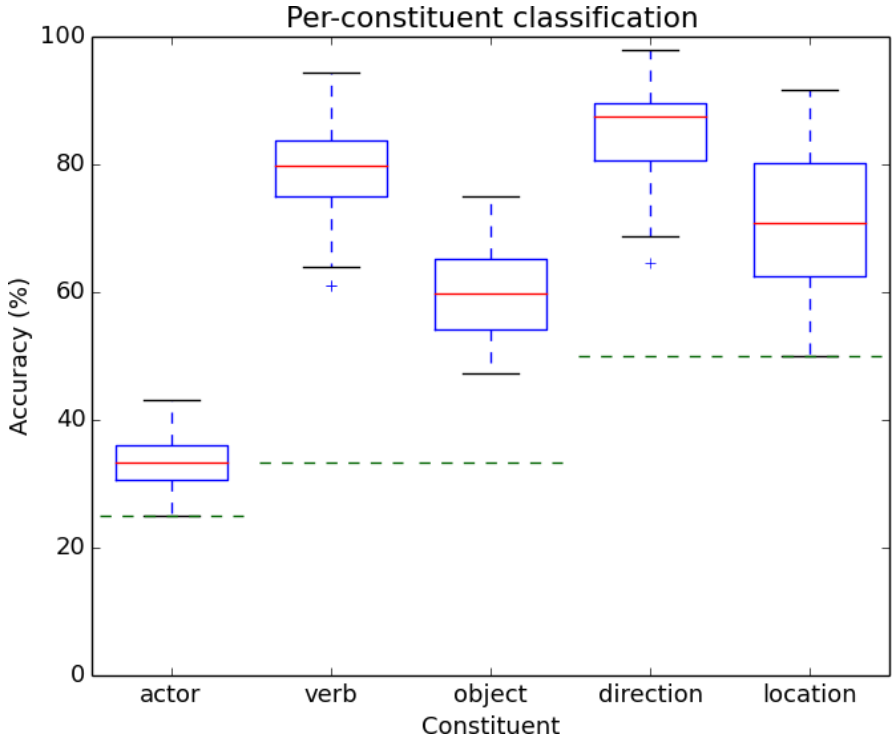


Figure 2: Results of per-constituent classification. (top) Per-subject mean classification accuracy for each constituent averaged across fold. Note that all five analyses perform above chance. (bottom) Classification accuracy aggregated across subject and fold for each constituent. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green lines indicates chance performance. A '*' indicates $p < 0.05$, a '**' indicates $p < 0.005$, and a '***' indicates $p < 0.0005$. (See the supplementary material for discussion.)

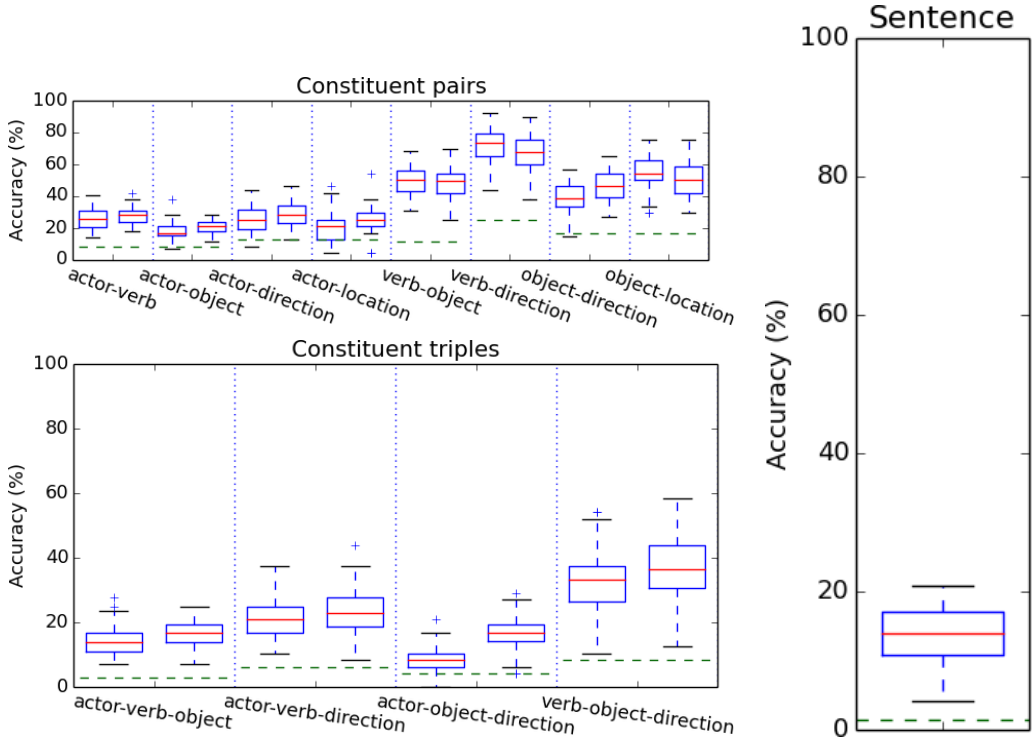


Figure 3: (a) Comparison of joint (left) *vs.* independent (right) classification accuracy aggregated across subject and fold for constituent pairs and triples. (b) Accuracy of classifying an entire sentence using independent per-constituent classifiers aggregated across subject and fold. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green lines indicates chance performance.

	actor verb	actor object	actor direction	actor location	verb object	verb direction	object direction	object location
accuracy	0.6607	0.7059	0.6336	0.6763	0.6709	0.7422	0.6544	0.6235
MCC	0.3724	0.3430	0.3603	0.2807	0.5959	0.7048	0.5560	0.5067

	actor verb object	actor verb direction	actor object direction	verb object direction
accuracy	0.8093	0.7403	0.8344	0.7154
MCC	0.2521	0.3004	0.2352	0.4594

Figure 4: Accuracy and MCC between independent and joint classification for constituent pairs (top) and triples (bottom), over the samples where the joint classifier was correct, aggregated across subject and fold.

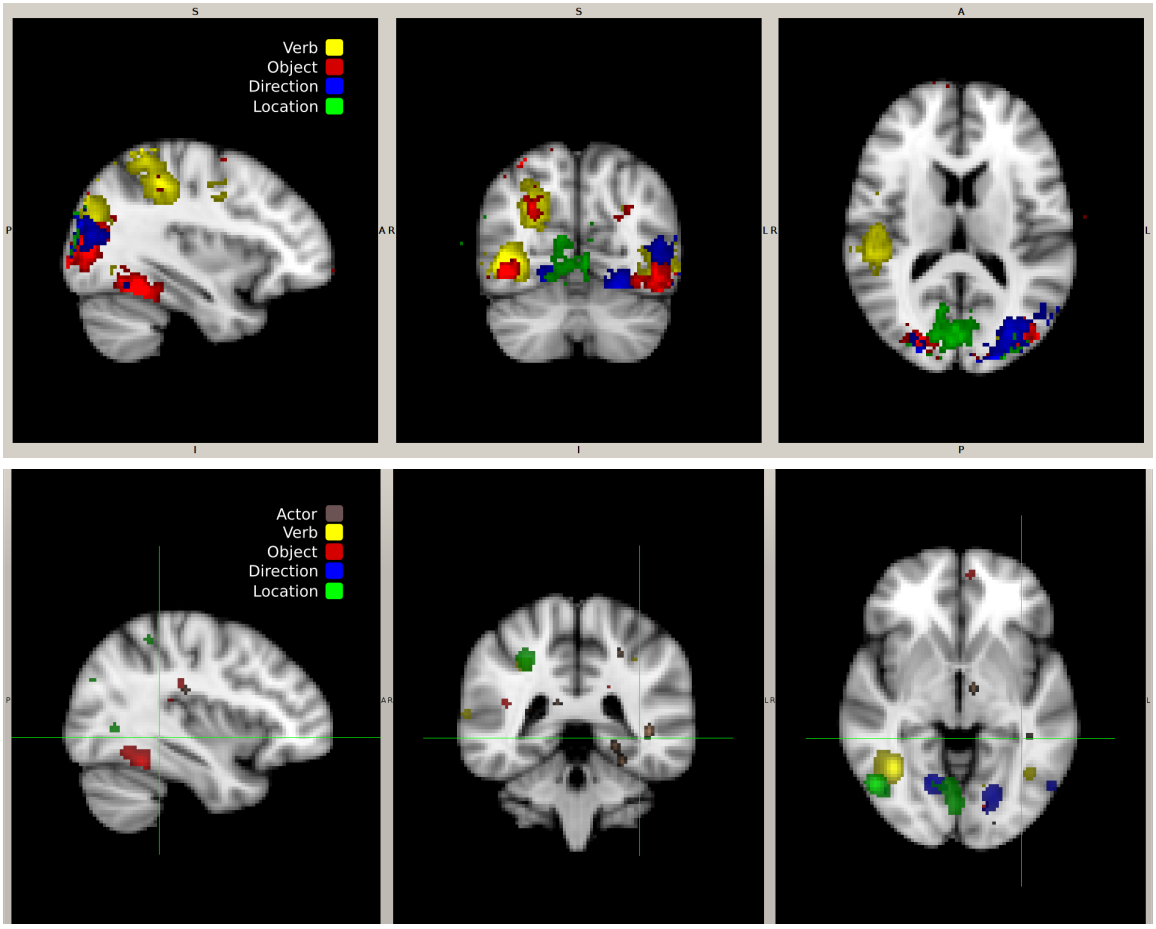


Figure 5: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 1 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 1, back-projected onto the anatomical scan, aggregated across run.

$$\frac{\left|\bigcap_i \text{independent}_i\right|}{\left|\bigcup_i \text{independent}_i\right|}$$

actor verb	actor object	actor direction	actor location	verb object	verb direction	object direction	object location
3.30%	6.74%	1.74%	1.32%	14.98%	1.20%	8.43%	4.48%
2.84%	2.54%	1.16%	2.06%	6.05%	3.61%	3.70%	2.36%

actor verb object	actor verb direction	actor object direction	verb object direction
1.26%	0.06%	0.65%	0.49%
0.42%	0.01%	0.00%	0.20%

$$\frac{\left|\left(\bigcup_i \text{independent}_i\right) \cap \text{joint}\right|}{|\text{joint}|}$$

actor verb	actor object	actor direction	actor location	verb object	verb direction	object direction	object location
67.42%	48.64%	68.53%	69.37%	79.53%	74.53%	65.97%	79.15%
58.85%	51.22%	42.42%	27.71%	66.05%	62.38%	52.70%	38.81%

actor verb object	actor verb direction	actor object direction	verb object direction
24.48%	59.43%	37.78%	55.13%
60.68%	56.51%	38.35%	58.25%

Figure 6: Quantitative comparison of the brain regions indicated by searchlight (upper rows) and thresholded SVM coefficients (lower rows) of the independent classifiers to the joint classifiers, for all constituent pairs and triples, averaged across subject. (top) The percentage of voxels in the union of the constituents for the independent classifier that are also in the intersection. (bottom) The percentage of voxels in the joint classifier that are shared with the independent classifier.

Supplementary Material for: The Compositional Nature of Event Representations in the Human Brain

Andrei Barbu,[1*] N. Siddharth,[1] Caiming Xiong,[2]
Jason J. Corso,[2] Christiane D. Fellbaum,[3]
Catherine Hanson,[4] Stephen José Hanson,[4]
Sébastien Hélie,[5] Evguenia Malaia,[6] Barak A. Pearlmutter,[7]
Jeffrey Mark Siskind,[1] Thomas Michael Talavage,[1] Ronnie B. Wilbur[8]
[1]School of Electrical and Computer Engineering
Purdue University, West Lafayette IN 47907
[2]Department of Computer Science and Engineering
SUNY Buffalo, Buffalo NY 14260
[3]Department of Computer Science, Princeton University, Princeton NJ 08540
[4]Department of Psychology and Rutgers Brain Imaging Center
Rutgers University, Newark NJ 07102
[5]Psychological Sciences, Purdue University, West Lafayette IN 47907
[6]Southwest Center for Mind, Brain, and Education
University of Texas at Arlington, Arlington TX 76019
[7]Hamilton Institute and Department of Computer Science
National University of Ireland Maynooth, Co. Kildare, Ireland
[8]Department of Speech, Language, and Hearing Sciences and Linguistics Program
Purdue University, West Lafayette IN 47907

[*]To whom correspondence should be addressed; E-mail: andrei@0xab.com.

Appendix to Fig. 1

Each stimulus depicted the combination of an actor, a verb, an object, and either a direction or a location. Direction was only depicted for the two verbs *carry* and *leave* while location was only depicted for the verb *fold*. There were four actors, three verbs, three objects, two directions, and two locations. This leads to $4 \times 3 \times 3 \times 2 = 72$ possible distinct depictions. The variation in direction and location was accomplished by mirroring videos about the vertical axis. All other variation was accomplished by filming a combination of actor, verb, and object. There were between 3 and 7 (mean 5.5) videos filmed for each combination. All stimulus videos were filmed in the same uncluttered uniform background which contained no other objects except for a table. The action depiction was intentionally varied to be unconventional (humorous) to keep subjects awake, attentive, and unhabituated. The runs were individually counterbalanced. Each run comprised 72 stimulus presentations, exactly one for each possible depiction. The particular film chosen for the depiction was randomly drawn from a uniform distribution. Some stimuli may have been chosen for multiple runs. Thus the training set contained exactly 7 times as many depictions for any combination of particular constituents as the test set, though the particular stimulus video for a given depiction may have appeared more than once in the training set and may have been shared between the training and test sets. The order of presentation depiction for each run was randomly drawn from a uniform distribution over permutations, but all subjects were presented with the same presentation order for runs and identical stimuli and stimulus order within the run. The video stimuli,

presentation files, and presentation software are available on the attached DVD. Scan data is available on request.

Appendix to Fig. 2

Figs. S1(a), S2(a), S3(a), S4(a), and S5(a) break down the results from Fig. 2 (bottom) by subject. Figs. S1(b), S2(b), S3(b), S4(b), and S5(b) present the corresponding confusion matrices, aggregated across subject and fold.

Appendix to Fig. 3

Fig. S6 (second and third portions) provides per-subject classification accuracy, including means and standard deviations across subjects, for constituent pairs and triples, averaged across fold, comparing joint classification (indicated with '-') to independent classification (indicated with '&'). Figs. S7 through S18 break down the results from Fig. 3(a) by subject. Fig. S19 breaks down the results from Fig. 3(b) by subject.

Appendix to Fig. 4

Fig. S20 extends Fig. 4 to include additional comparisons between the independent and joint classifiers.

Appendix to Fig. 5

Figs. S21 through S27 extend Fig. 5 with analyses for all subjects.

Appendix to Fig. 6

Figs. S28 and S29 break down the results from Fig. 6 by subject.

Analyses

Fig. S30 reports the training- and test-set sizes for each of our classification analyses. We conducted 30 classification analyses in total: 5 single-constituent analyses, 8 constituent-pair analyses, both independent and joint, 4 constituent-triple analyses, both independent and joint, and an independent sentence analysis. The differences in the various sizes result from several properties. First, verb does not combine with location since location only applies to a single verb, *fold*. Second, we did not train a joint classifier for actor, object, and location because we would have only 7 training samples per subject, fold, and class. Similarly, we employed only an independent classifier for sentence and did not train a joint classifier because we would have only 7 training samples per subject, fold, and class.

Statistical Significance

For Figs. 2 (top) and S6, and references thereof in the text, we indicate $p < 0.05$ with a '*', $p < 0.005$ with a '**', and $p < 0.0005$ with a '***'. As per Fig. S30, per-subject classification results are over 192 trials for analyses that involve location, 384 trials for analyses that involve direction, and 576 trials for all other analyses. Classification results aggregated across subjects are over 1344 trials for analyses that involve location, 2688 trials for analyses that involve direction, and 4032 trials for all other analyses. We used a χ^2 test to compute p values for all classification results. In most cases, this leads to extremely small values, because the null hypothesis is a binomial distribution comprised of repeated independent Bernoulli trials with a uniform distribution over possible outcomes. Assuming independence between trials where each trial is uniformly distributed is warranted because all runs were counterbalanced with random

presentation order drawn from a uniform distribution over permutations, filmed in the same uncluttered uniform background, and the action depiction was intentionally varied to be unconventional (humorous) to keep subjects awake, attentive, and unhabituated. We report 30 aggregate analyses across subject. All are highly significant; the largest p value was less than 10^{-21}. We omit '***' annotations for Figs. 2 (bottom), 3, S1 through S5, and S7 through S19 as all such results aggregate analyses across subject. We report 210 per-subject analyses. Of these, only 14 instances have p values that exceed 0.05: the single-constituent actor analysis for subjects 1 and 2, the joint actor-location analysis for subjects 1, 2, 4, 5, and 7, the independent actor-location analysis for subjects 1, 2, and 3, the independent actor-verb-direction analysis for subject 3, and the joint actor-object-direction analysis for subjects 2, 4, and 7. These apply to Figs. 2 (top) and S6, and references thereof in the text. We know of no way to determine statistical significance of the results in Figs. 4, 6, S20, S28 and S29.

Independence

Note that we are claiming that the brain independently processes constituents, *e.g.*, verb and object, *not* that the output of such processing is independent. In particular, we are *not* claiming that the output of our classifiers are independent across constituent. Classification results are produced by a long pipeline: the stimulus, the evoked brain activity, its indirect measurement via fMRI, and its analysis via classification. Cross-constituent dependence can be introduced at any stage in this pipeline and could also be masked by any subsequent stage. Moreover, our classifiers are imperfect. The confusion matrices are not diagonal. Since our design is counterbalanced, in order for a χ^2 test not to reject the null hypothesis, the contingency table must be uniform. However, if the verb classifier exhibits a misclassification bias where, for example, *carry* is misclassified as *fold* more frequently than as *leave*, and the object classifier exhibits a similar misclassification bias, where, for example, *chair* is misclassified as *shirt* more frequently than as *tortilla*, this would manifest as dependence between verb and object in the classifier output that would have no bearing on classification accuracy. Nor would it indicate joint usage of verb and object information during classification. Thus it makes no sense to perform a standard χ^2 independence test between constituent pairs of classifier outputs.

What we are claiming is that the brain largely makes classification decisions for one constituent independent of those for other constituents. We take as evidence for this the fact that

- Classification accuracy using independent classifiers is largely the same as that for corresponding joint classifiers (Figs. 3, 4 and S20).
- The brain regions employed by the per-constituent classifiers are largely pairwise disjoint (Figs. 6 and S28).
- The brain regions employed by the joint classifiers largely consist of the unions of the brain regions employed by the component constituent classifiers (Figs. 6 and S28).

Linguistic vs. Visual Stimuli and Tasks

Our stimuli were purely visual. There were no words, phrases, or sentences presented, either auditorily or visually (orthographic, signed). Our subjects were given no specific task. No behavioral or motor response of any kind was elicited. Specifically, subjects were not asked to produce words, phrases, or sentences, either oral or visual (written, signed). Thus neither the stimuli nor the (nonexistent) task were overtly linguistic. Nonetheless, the experimental setup was implicitly linguistic in a number of ways. Subjects were shown sample video prior to imaging and were informed of the structure of the stimuli and the intended collection of actors, verbs, objects, directions, and locations. All subjects were aware of the experiment design, were informed of the intended depiction of each stimulus prior to the scan, and were instructed to think of the intended depiction after each presentation. While they had no overt task, they were asked to think about the sentence depicted by each video. It is conceivable that such subject instruction introduced a linguistic aspect to the task and is what induced a decomposable representation. Perhaps the subjects were internally vocalizing the sentences that they were thinking?

This would be interesting in its own right, as it would indicate generation of internal linguistic representations even given a lack of overt linguistic behavioral and motor response. Nonetheless, it would be interesting to see if such

representations arose even when subjects were not given such explicit instructions and perhaps were not even primed as to the experiment design, the set of target constituents, and the set of classes within each constituent. Moreover, it would be interesting to see if such representations also arise for stimuli that are less conducive to sentential description, such as more abstract perhaps synthetic video of moving shapes that nonetheless could be conceptually decomposed into shape *vs.* motion patterns *vs.* direction and location that wouldn't be described as nouns, verbs, and prepositions.

Jackendoff's Cognitive Constraint and Conceptual Structure Hypothesis

The *Cognitive Constraint* and *Conceptual Structure Hypothesis* are discussed by Jackendoff (*1*).

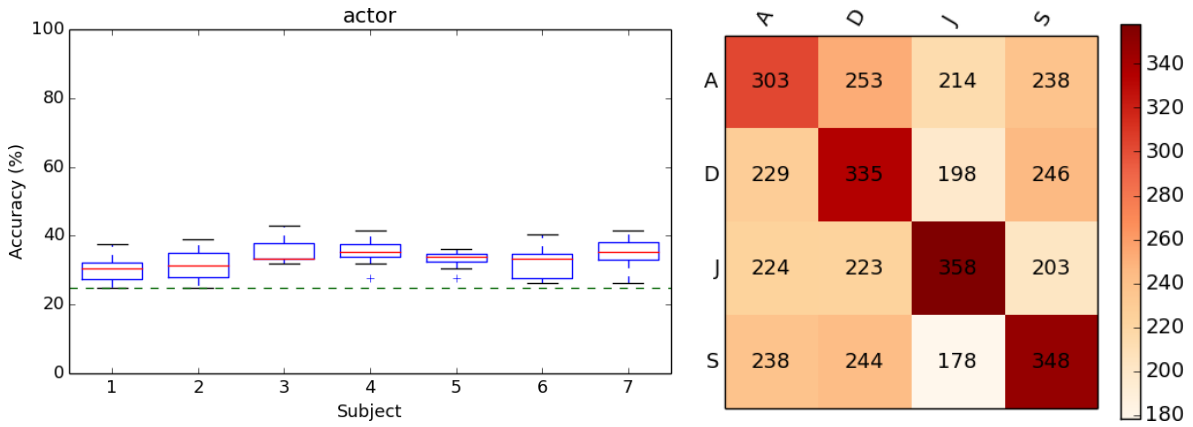
> I will call this constraint the Cognitive Constraint: There must be levels of mental representation at which information conveyed by language is compatible with information from other peripheral systems such as vision, nonverbal audition, smell, kinesthesia, and so forth. If there were no such levels, it would be impossible to use language to report sensory input. We couldn't talk about what we see and hear. Likewise, there must be a level at which linguistic information is compatible with information eventually conveyed to the motor system, in order to account for our ability to carry out orders and instructions.
>
> (p. 16)

> The Conceptual Structure Hypothesis
>
> There is a single level of mental representation, conceptual structure, at which linguistic, sensory, and motor information are compatible.
>
> (p. 17)

References

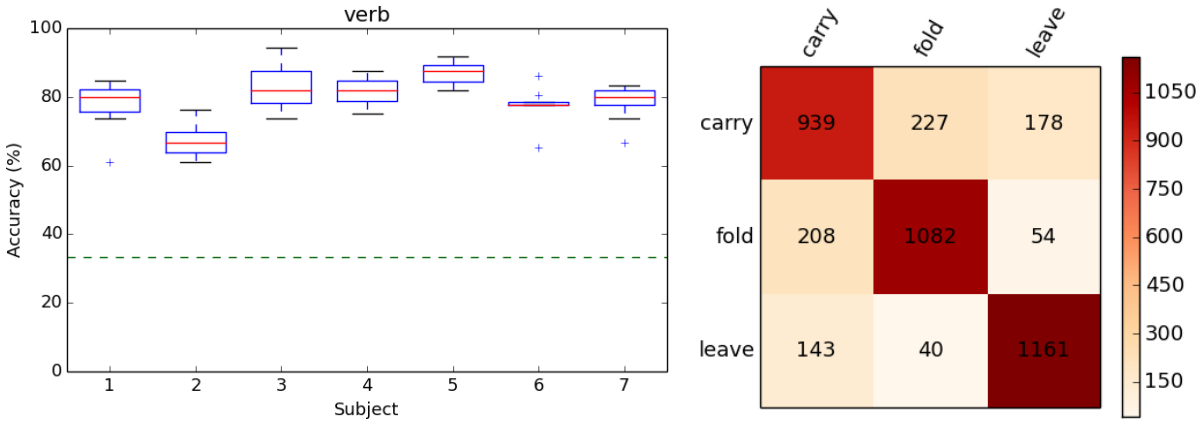
1. R. Jackendoff, Semantics and cognition (1983). See pp. 16–22 including (1.3) and (1.4).

2. R. Jackendoff, *Semantics and Cognition* (MIT Press, Cambridge, MA, 1983).



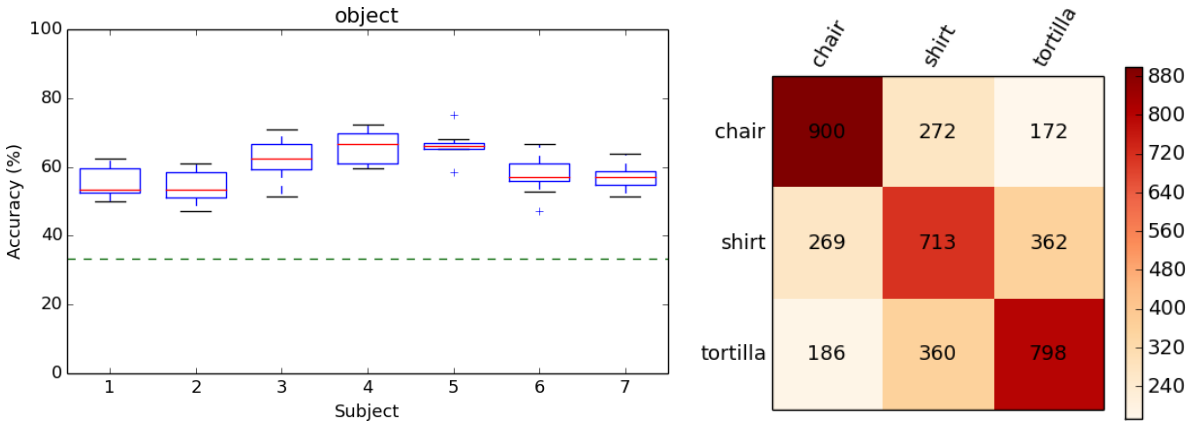
(a) (b)

Figure S1: (a) Per-subject classification accuracy for **actor** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance. (b) Corresponding confusion matrix, aggregated across subject and fold. Note that it is largely diagonal.



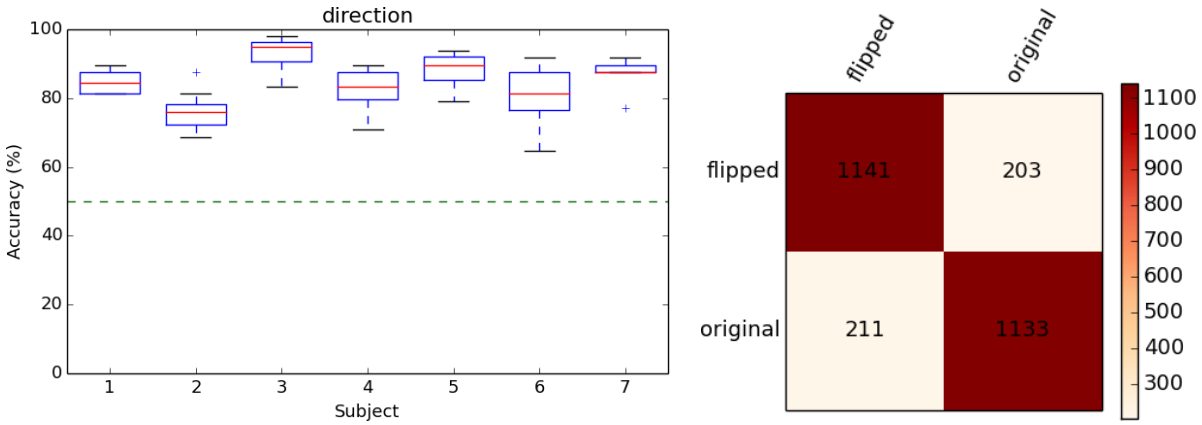
(a) (b)

Figure S2: (a) Per-subject classification accuracy for **verb** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance. (b) Corresponding confusion matrix, aggregated across subject and fold. Note that it is largely diagonal.



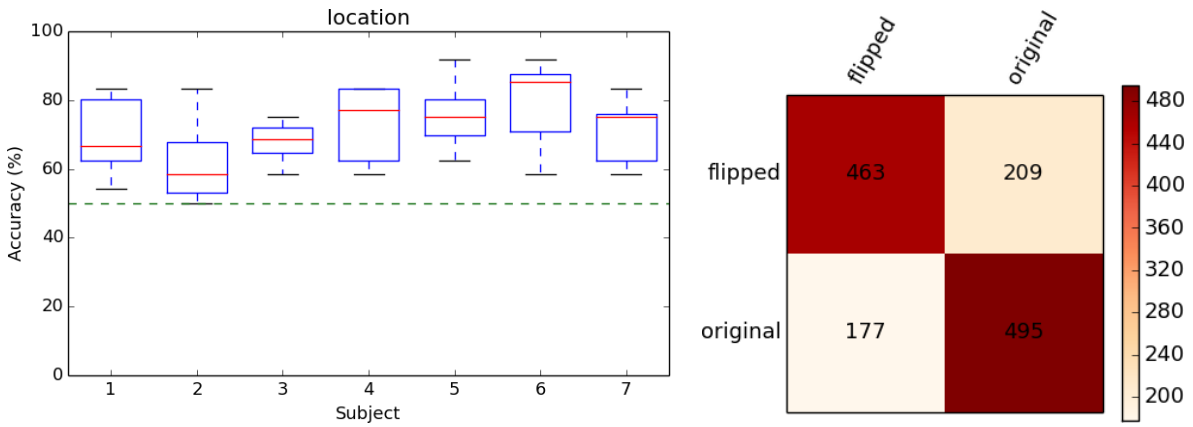
(a) (b)

Figure S3: (a) Per-subject classification accuracy for **object** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance. (b) Corresponding confusion matrix, aggregated across subject and fold. Note that it is largely diagonal.



(a) (b)

Figure S4: (a) Per-subject classification accuracy for **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance. (b) Corresponding confusion matrix, aggregated across subject and fold. Note that it is largely diagonal.



(a) (b)

Figure S5: (a) Per-subject classification accuracy for **location** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance. (b) Corresponding confusion matrix, aggregated across subject and fold. Note that it is largely diagonal.

analysis	chance	mean	stddev	1	2	3	4	5	6	7
actor	0.2500	0.3333***	0.045	0.304	0.314	0.356***	0.352***	0.332**	0.323**	0.352***
verb	0.3333	0.7892***	0.079	0.776***	0.674***	0.830***	0.816***	0.870***	0.776***	0.783***
object	0.3333	0.5980***	0.067	0.554***	0.542***	0.623***	0.660***	0.663***	0.575***	0.569***
direction	0.5000	0.8460***	0.076	0.846***	0.763***	0.930***	0.823***	0.880***	0.807***	0.872***
location	0.5000	0.7128***	0.110	0.698***	0.615*	0.677***	0.734***	0.755***	0.797***	0.714***
actor-verb	0.0833	0.2579***	0.063	0.210***	0.214***	0.302***	0.264***	0.292***	0.259***	0.266***
actor&verb	0.0833	0.2686***	0.054	0.236***	0.220***	0.295***	0.293***	0.295***	0.252***	0.288***
actor-object	0.0833	0.1756***	0.055	0.123*	0.158***	0.193***	0.220***	0.184***	0.175*	0.175***
actor&object	0.0833	0.2061***	0.041	0.167***	0.170***	0.224***	0.229***	0.233***	0.208***	0.212***
actor-direction	0.1250	0.2504***	0.084	0.198***	0.195**	0.302***	0.263***	0.227***	0.273***	0.294***
actor&direction	0.1250	0.2846***	0.071	0.260***	0.258***	0.313***	0.323***	0.289***	0.234***	0.315***
actor-location	0.1250	0.2031***	0.095	0.161	0.177	0.245***	0.141	0.203	0.297***	0.198
actor&location	0.1250	0.2403***	0.079	0.208	0.182	0.240	0.224*	0.245***	0.302***	0.281***
verb-object	0.1111	0.4958***	0.092	0.523***	0.389***	0.540***	0.545***	0.595***	0.462***	0.417***
verb&object	0.1111	0.4794***	0.089	0.439***	0.366***	0.514***	0.547***	0.589***	0.437***	0.464***
verb-direction	0.2500	0.7143***	0.111	0.737***	0.581***	0.766***	0.727***	0.828***	0.596***	0.766***
verb&direction	0.2500	0.6711***	0.115	0.661***	0.505***	0.784***	0.682***	0.766***	0.625***	0.674***
object-direction	0.1667	0.3906***	0.094	0.354***	0.276***	0.456***	0.453***	0.471***	0.365***	0.359***
object&direction	0.1667	0.4621***	0.100	0.427***	0.346***	0.544***	0.513***	0.542***	0.414***	0.448***
object-location	0.1667	0.5513***	0.107	0.604***	0.557***	0.599***	0.505***	0.536***	0.599***	0.458***
object&location	0.1667	0.5000***	0.110	0.453***	0.437***	0.469***	0.557***	0.563***	0.552***	0.469***
actor-verb-object	0.0278	0.1434***	0.045	0.125***	0.099***	0.179***	0.177***	0.161***	0.149***	0.113***
actor&verb&object	0.0278	0.1687***	0.042	0.135***	0.123***	0.184***	0.193***	0.210***	0.161***	0.174***
actor-verb-direction	0.0625	0.2139***	0.061	0.188***	0.172***	0.279***	0.203***	0.253***	0.180***	0.224***
actor&verb&direction	0.0625	0.2333***	0.075	0.201***	0.182***	0.271	0.284***	0.260***	0.185***	0.250***
actor-object-direction	0.0417	0.0867***	0.040	0.083*	0.068	0.117***	0.060	0.096***	0.099*	0.083
actor&object&direction	0.0417	0.1633***	0.055	0.138***	0.107***	0.185***	0.206***	0.198***	0.138***	0.172***
verb-object-direction	0.0833	0.3255***	0.102	0.375***	0.182***	0.318***	0.401***	0.417***	0.289***	0.297***
verb&object&direction	0.0833	0.3679***	0.110	0.339***	0.227***	0.445***	0.430***	0.474***	0.305***	0.357***
sentence&	0.0139	0.1384***	0.043	0.113***	0.087***	0.149***	0.167***	0.168***	0.135***	0.149***

Figure S6: Per-subject classification accuracy, including means and standard deviations across subjects, for different classifiers, averaged across fold. Joint classifiers are indicated with '-'. Independent classifiers are indicated with '&'.

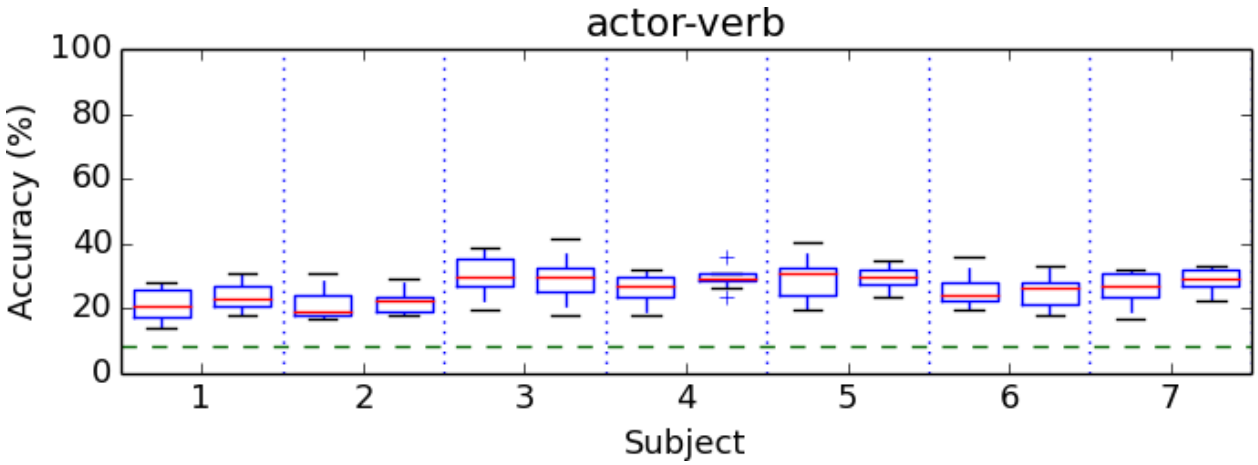


Figure S7: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor** and **verb** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

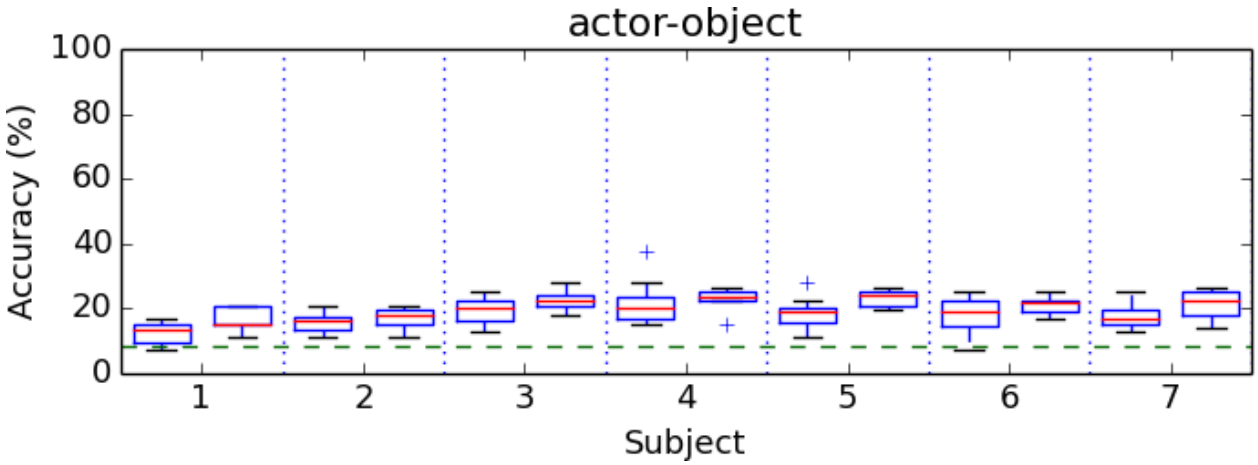


Figure S8: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor** and **object** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

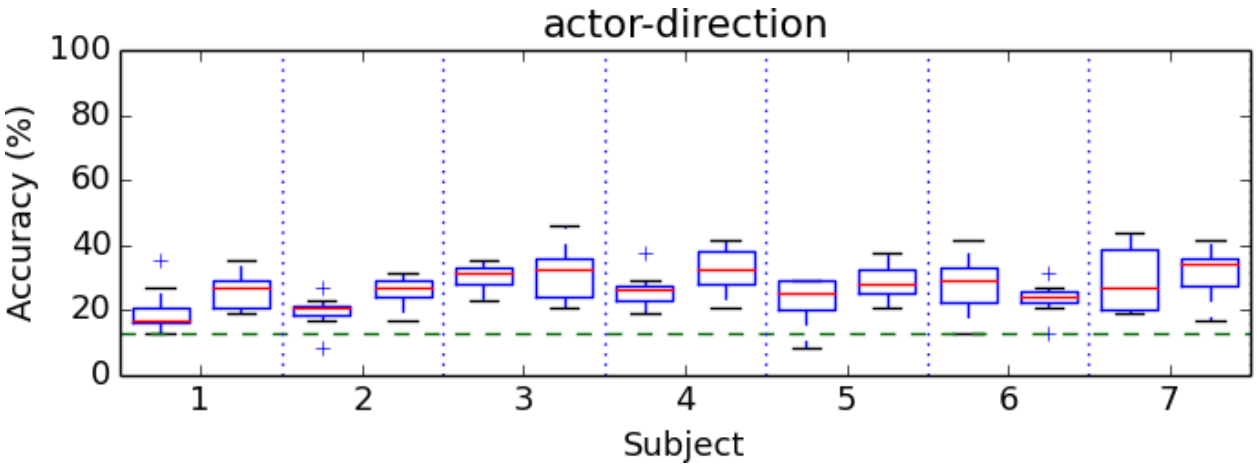


Figure S9: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor** and **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

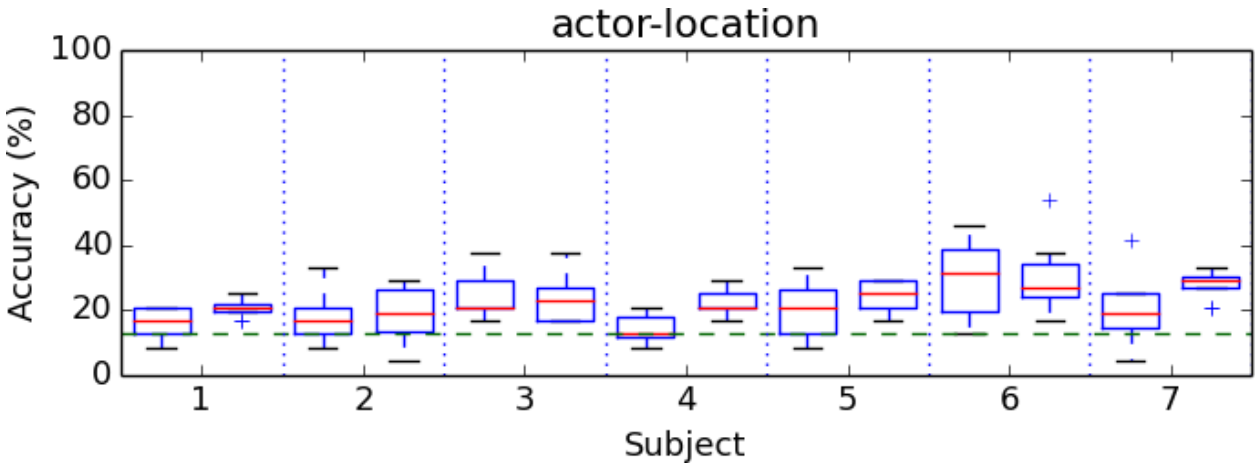


Figure S10: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor** and **location** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

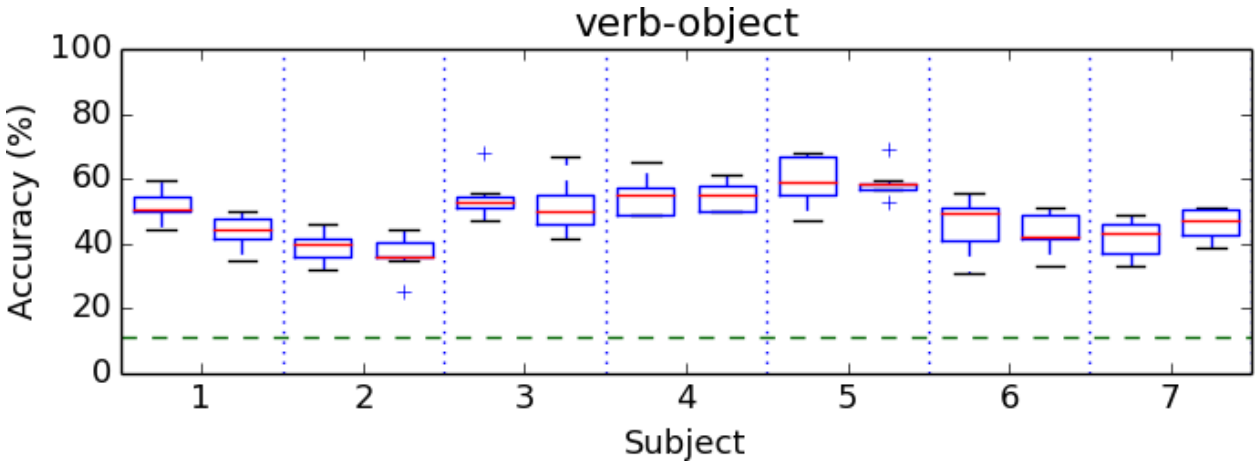


Figure S11: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **verb** and **object** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

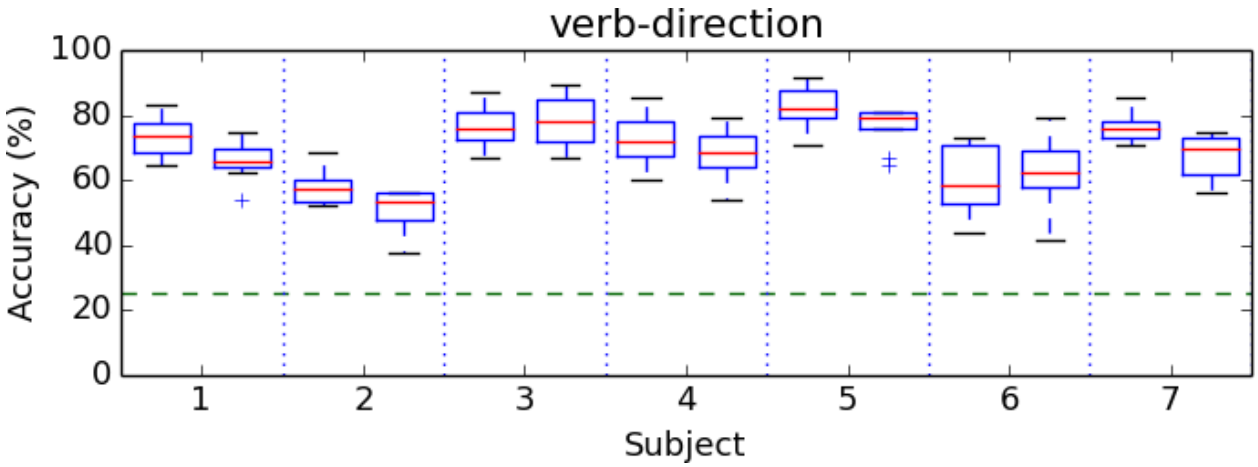


Figure S12: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **verb** and **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

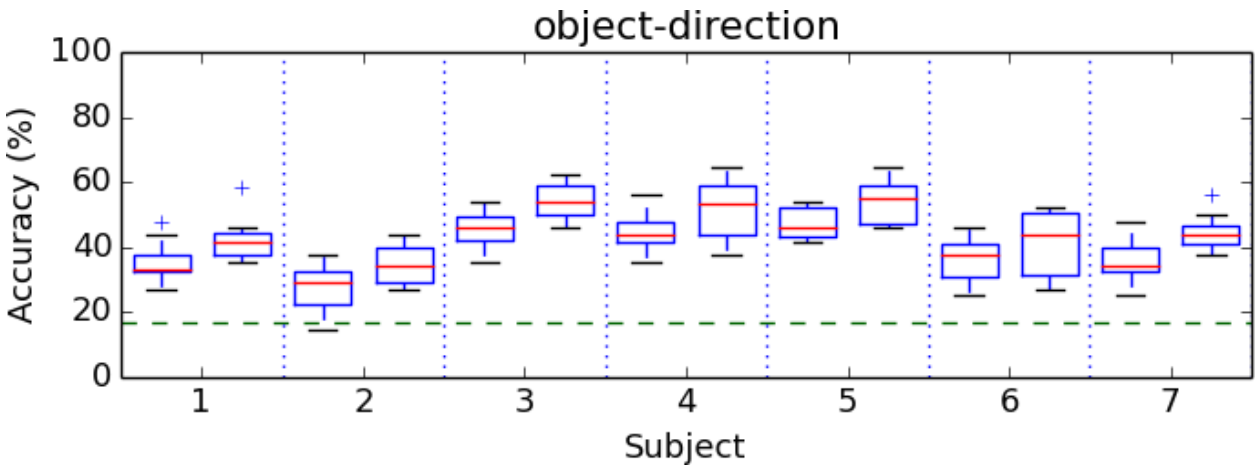


Figure S13: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **object** and **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

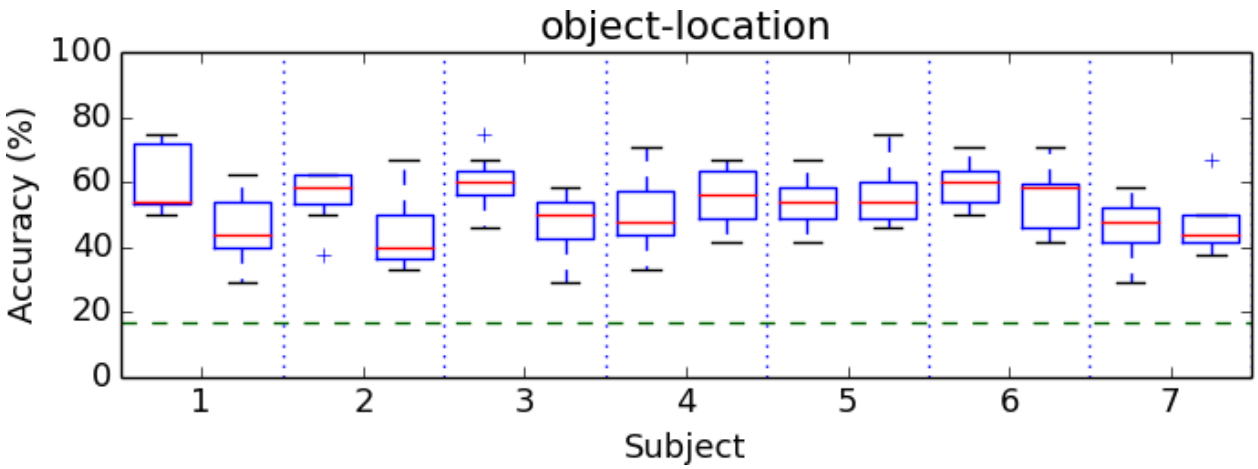


Figure S14: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **object** and **location** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

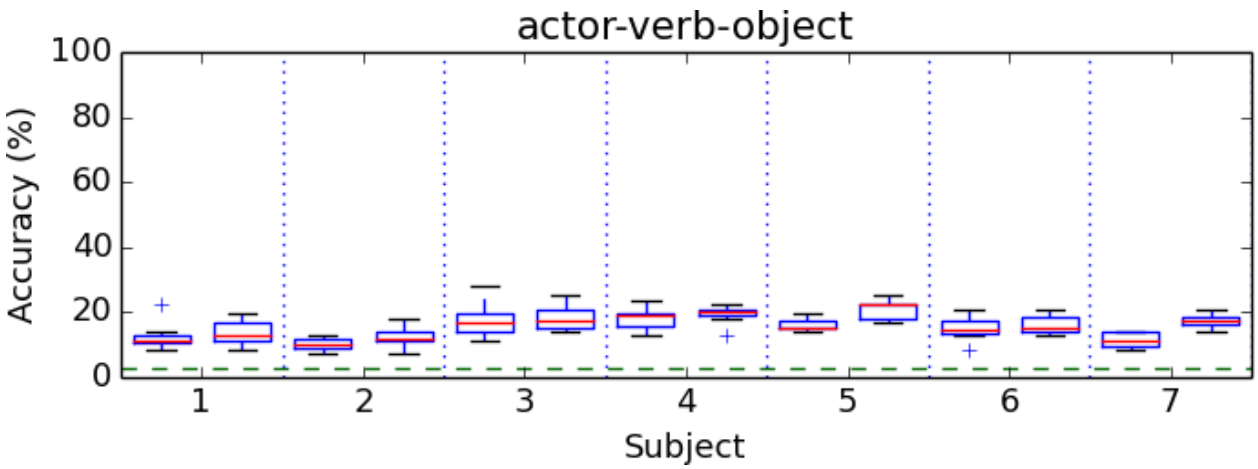


Figure S15: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor**, **verb**, and **object** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

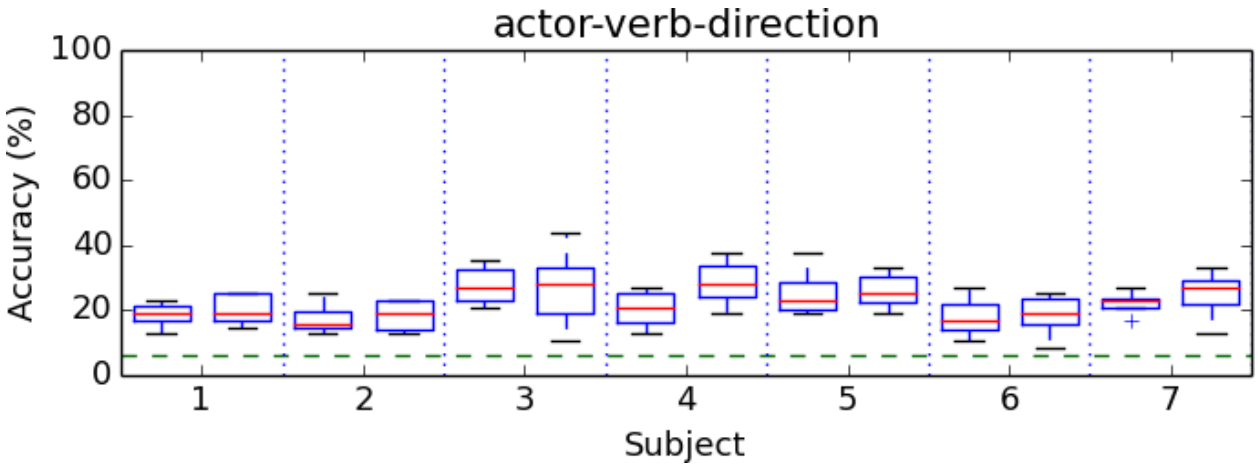


Figure S16: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor**, **verb**, and **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

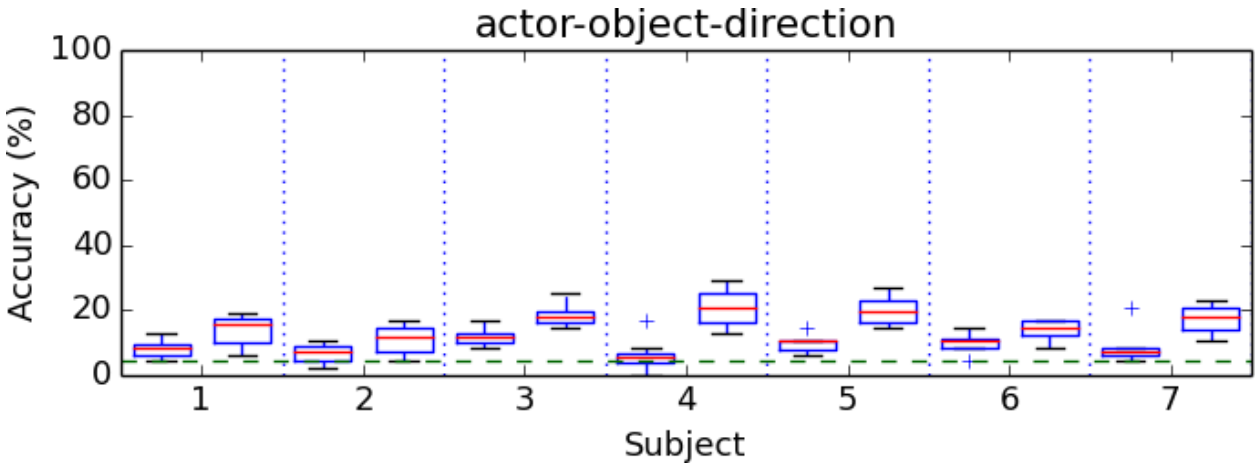


Figure S17: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **actor**, **object**, and **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

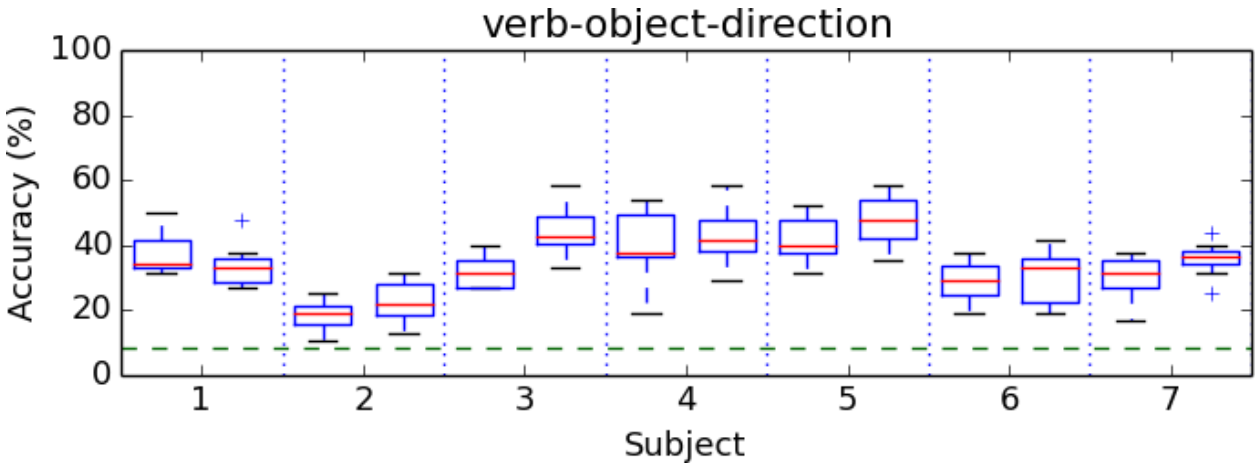


Figure S18: Per-subject comparison of joint (left) *vs.* independent (right) classification accuracy for the combination of **verb**, **object**, and **direction** across the different folds. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

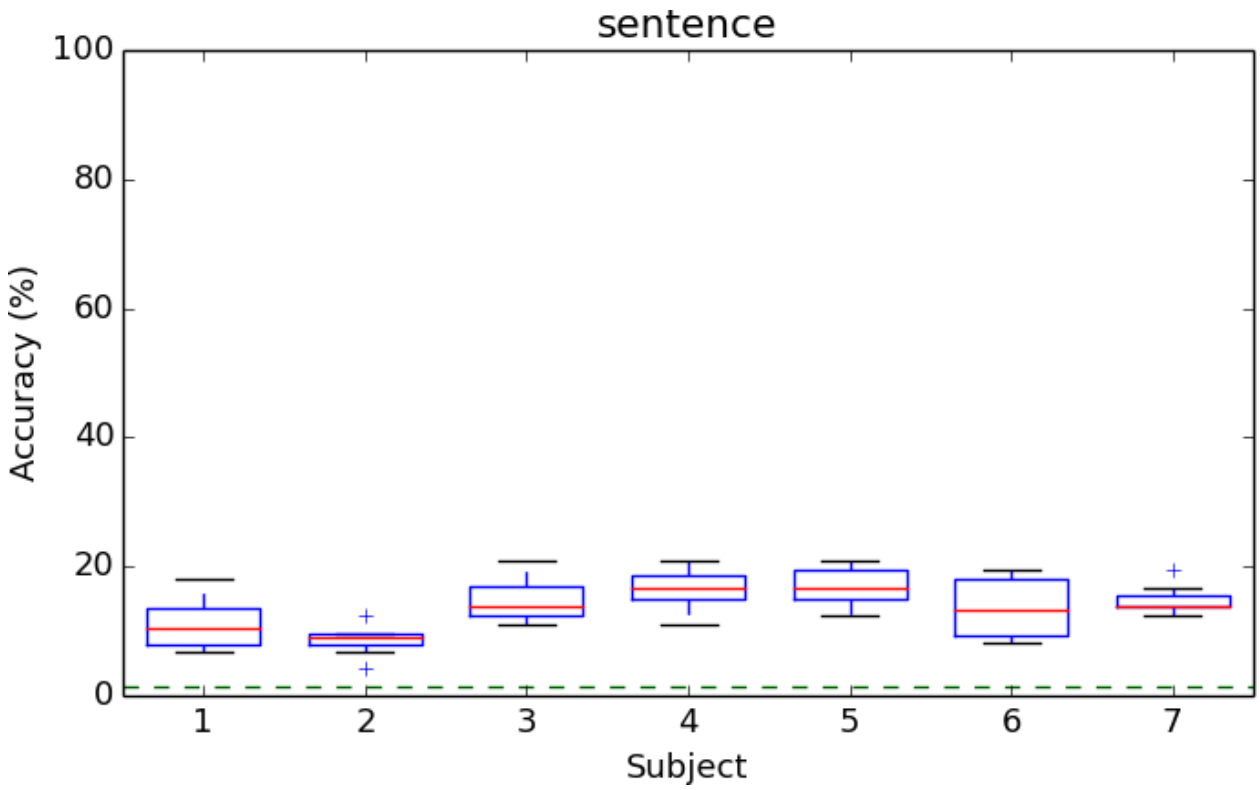


Figure S19: Per-subject classification accuracy for an entire sentence using independent per-constituent classifiers. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green line indicates chance performance.

analysis	acc-bit	mcc-bit	acc-all	mcc-all	acc-good	mcc-good
actor-verb	0.6607	0.1600	0.3006	0.2371	0.4250	0.3724
actor-object	0.7059	0.1475	0.2584	0.1910	0.3983	0.3430
actor-direction	0.6336	0.1266	0.3363	0.2412	0.4398	0.3603
actor-location	0.6763	0.1149	0.2723	0.1665	0.3736	0.2807
verb-object	0.6709	0.3419	0.4712	0.4051	0.6433	0.5959
verb-direction	0.7422	0.4172	0.6440	0.5433	0.7703	0.7048
object-direction	0.6544	0.2946	0.4475	0.3370	0.6305	0.5560
object-location	0.6235	0.2537	0.4754	0.3702	0.5897	0.5067
actor-verb-object	0.8093	0.1833	0.1262	0.1010	0.2751	0.2521
actor-verb-direction	0.7403	0.1979	0.1916	0.1432	0.3409	0.3004
actor-object-direction	0.8344	0.1314	0.1250	0.0867	0.2661	0.2352
verb-object-direction	0.7154	0.3318	0.2850	0.2267	0.5006	0.4594

Figure S20: Comparison of independent classifiers with joint classifiers, aggregated across subject and fold. 'Acc' denotes accuracy and 'mcc' denotes Matthews correlation coefficient (MCC). The 'bit' values involve computing a binary correct/incorrect label for each sample with both the independent and joint classifiers and computing the accuracy and MCC over the samples between the independent and joint classifiers. The 'all' values involve computing a (nonbinary) class label for each sample with both the independent and joint classifiers and computing the accuracy and MCC over the samples between the independent and joint classifiers. The 'good values' involved computing accuracy and MCC over the samples between the independent and joint classifiers for only those 'all' samples where the joint classifier is correct.

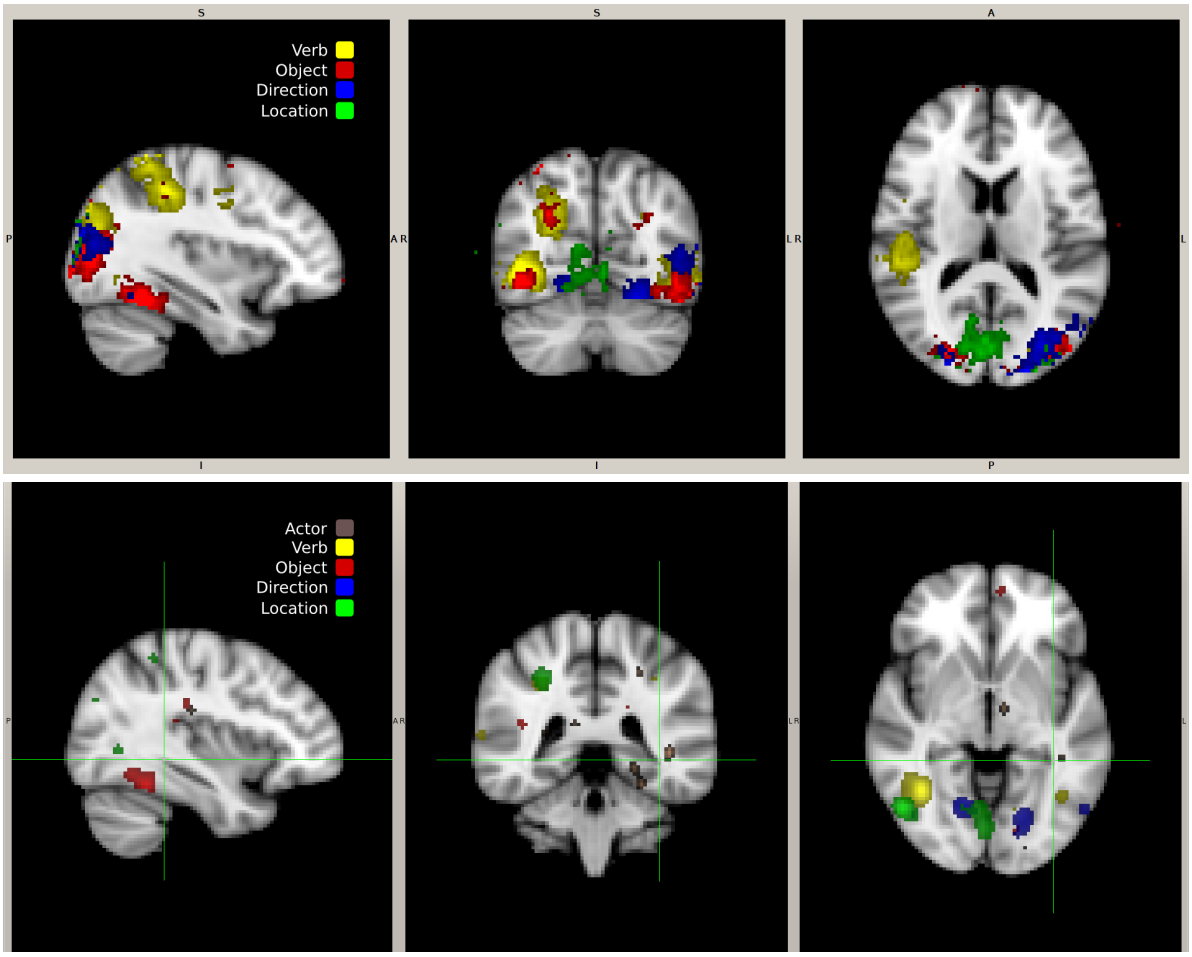


Figure S21: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 1 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 1, back-projected onto the anatomical scan, aggregated across run.

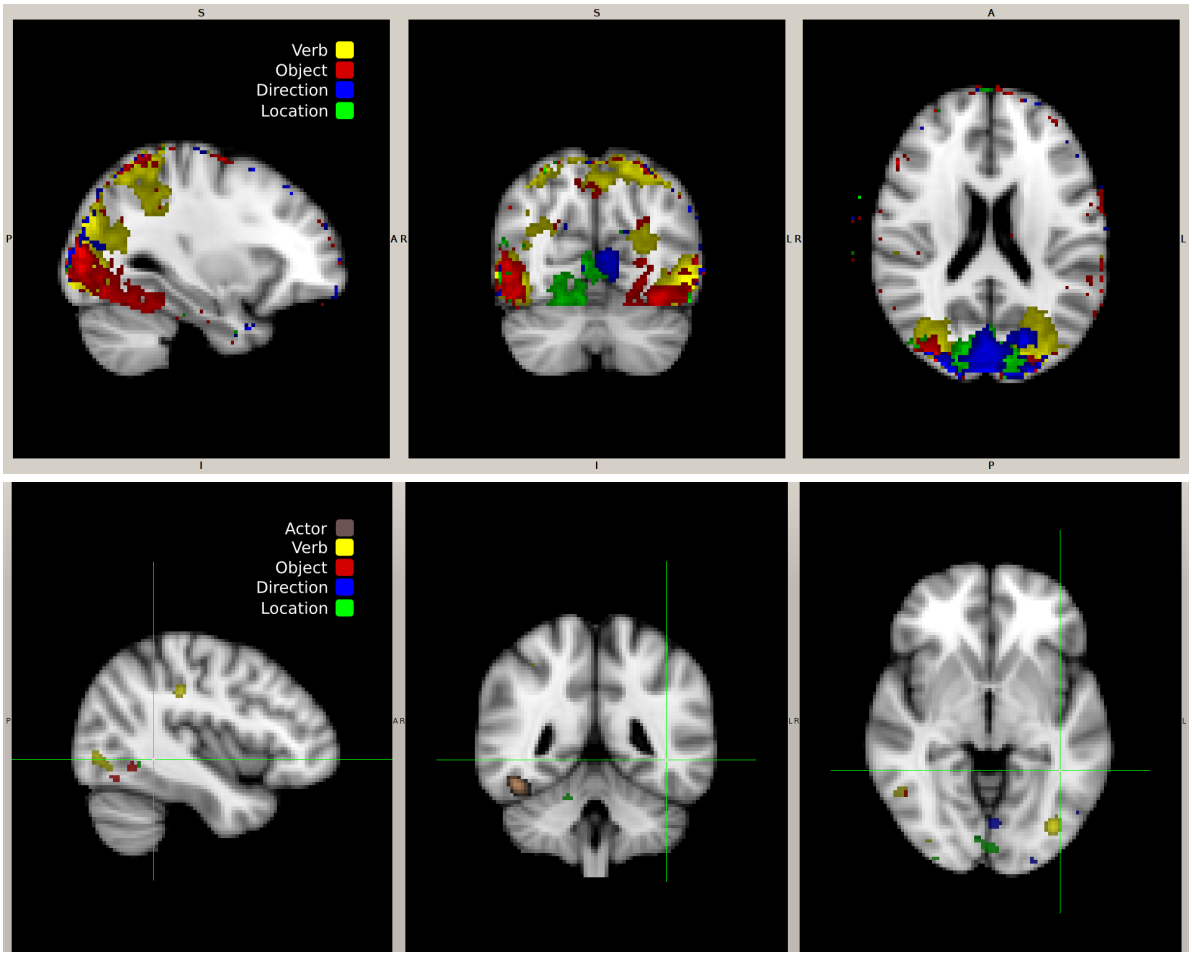


Figure S22: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 2 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 2, back-projected onto the anatomical scan, aggregated across run.

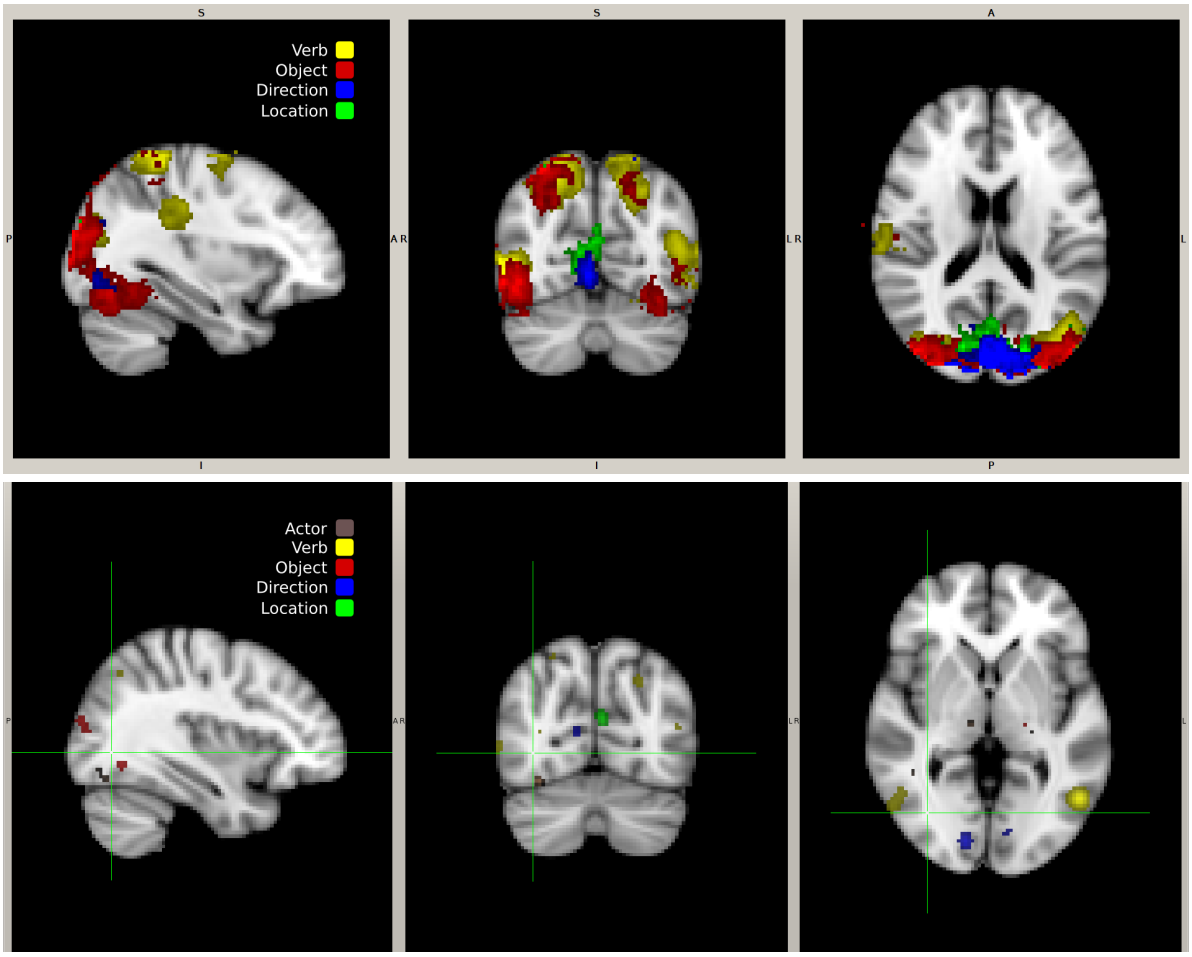


Figure S23: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 3 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 3, back-projected onto the anatomical scan, aggregated across run.

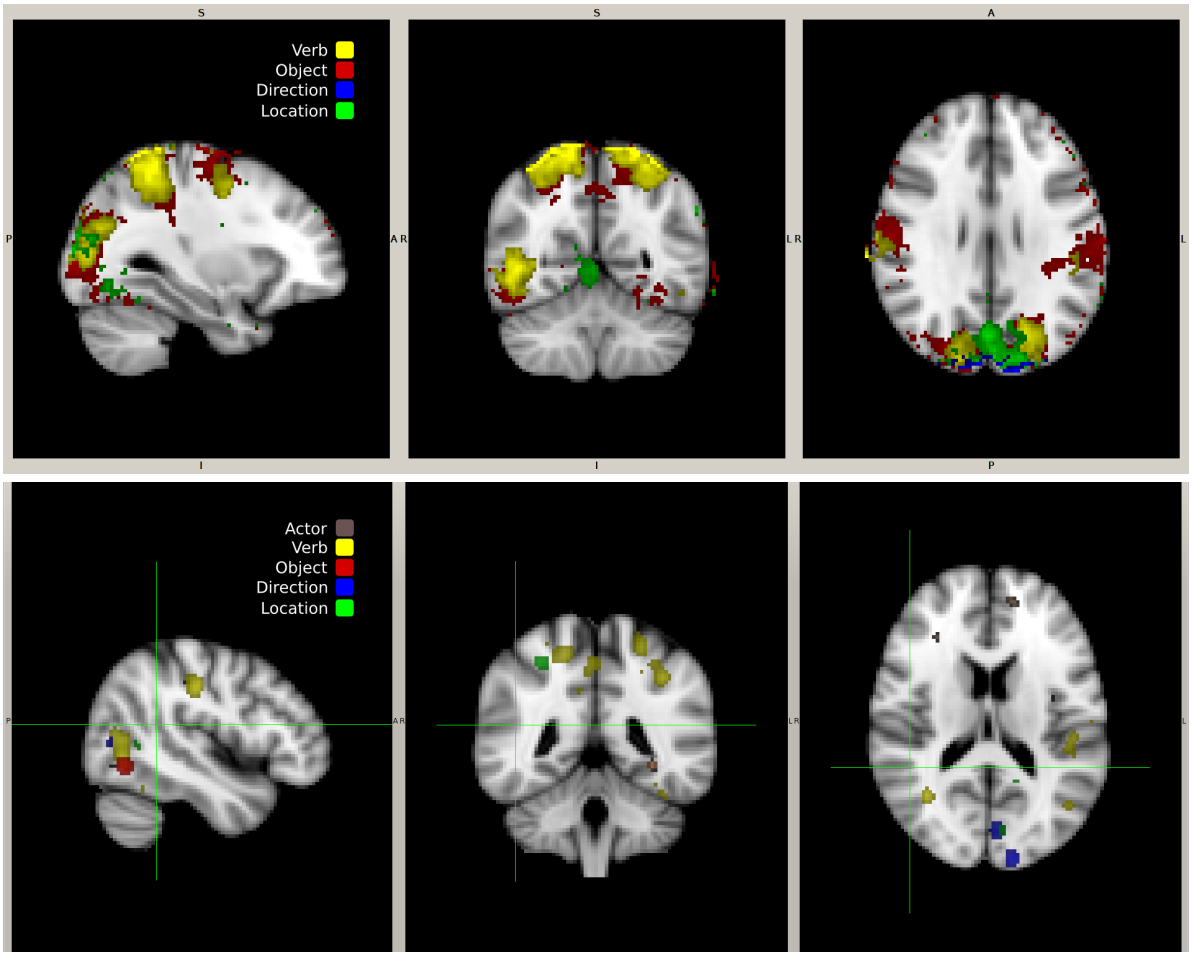


Figure S24: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 4 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 4, back-projected onto the anatomical scan, aggregated across run.

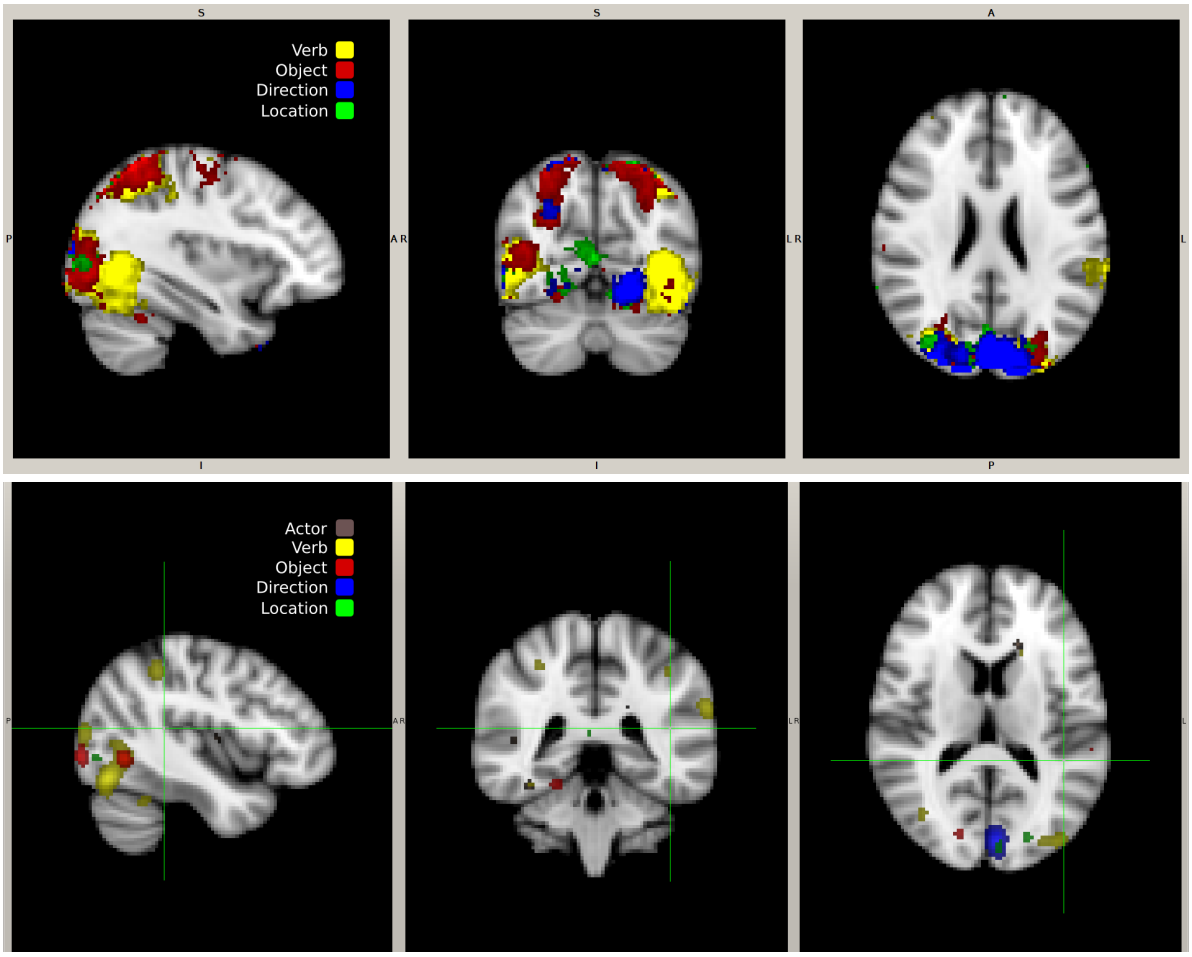


Figure S25: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 5 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 5, back-projected onto the anatomical scan, aggregated across run.

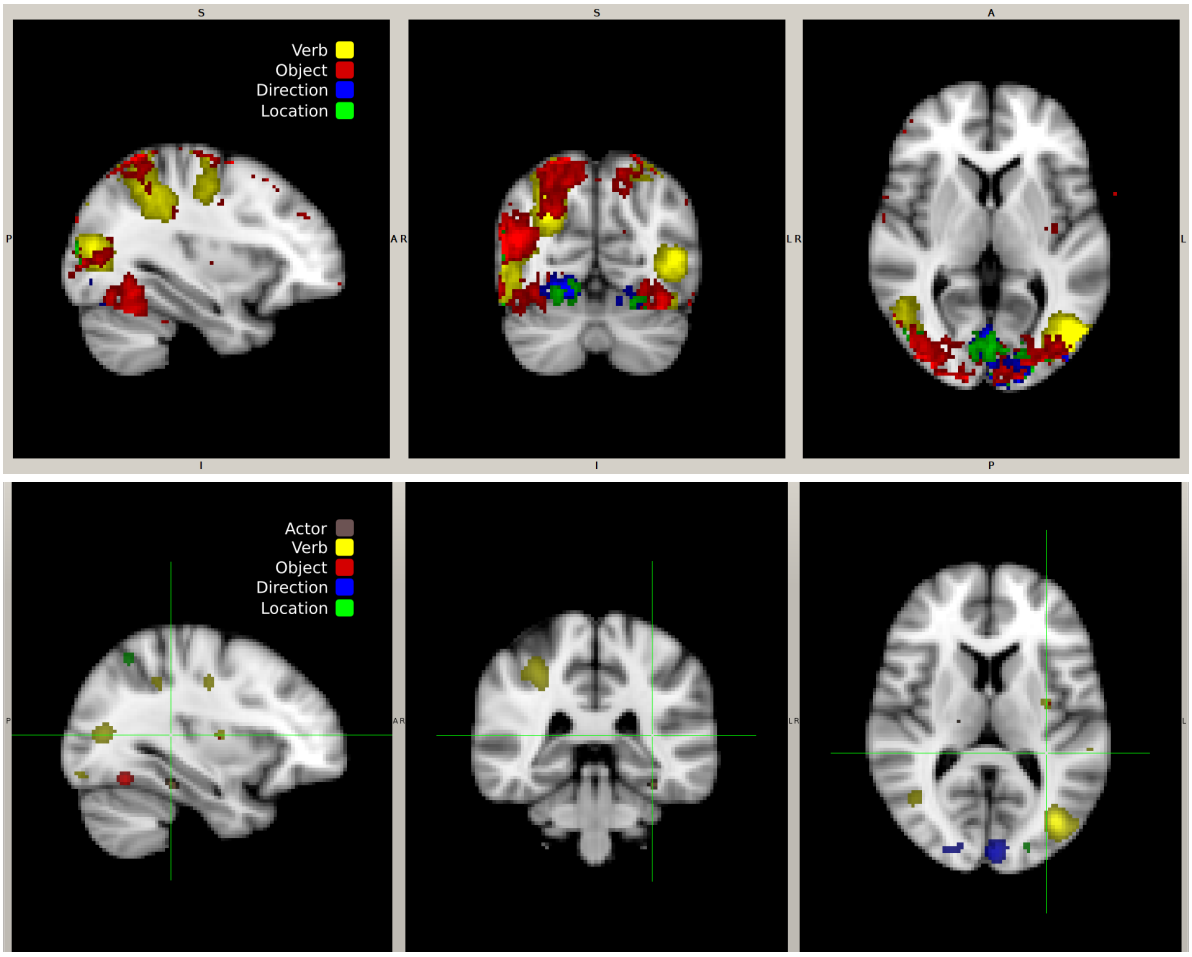


Figure S26: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 6 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 6, back-projected onto the anatomical scan, aggregated across run.

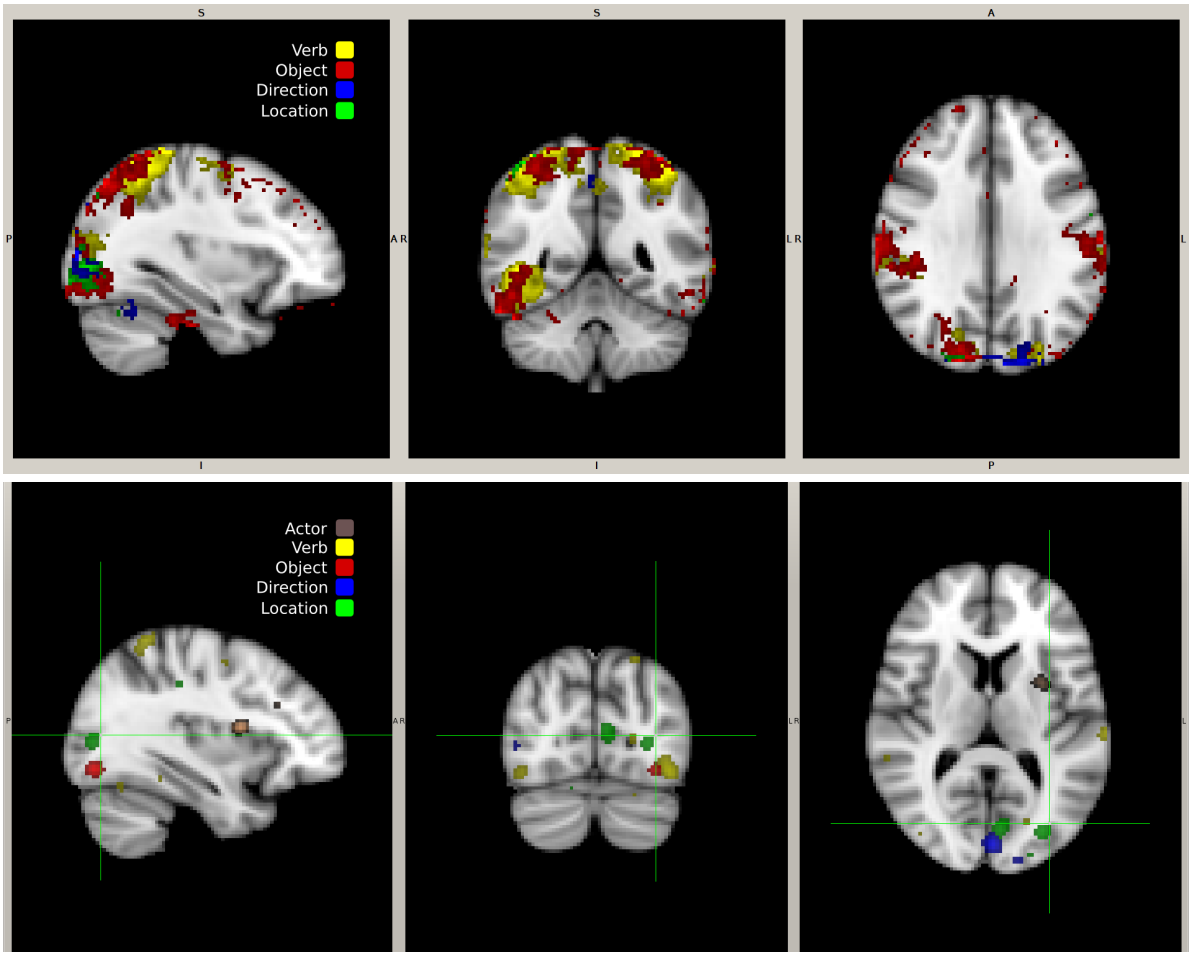


Figure S27: (top) Searchlight analysis indicating the classification accuracy of different brain regions on the anatomical scans from subject 7 averaged across stimulus, class, and run. (bottom) Thresholded SVM coefficients for subject 7, back-projected onto the anatomical scan, aggregated across run.

$$\frac{\left|\bigcap_i \text{independent}_i\right|}{\left|\bigcup_i \text{independent}_i\right|}$$

analysis	1	2	3	4	5	6	7	mean
actor-verb	0.78%	5.75%	3.55%	3.39%	4.44%	2.57%	2.63%	3.30%
actor-object	3.72%	4.73%	14.46%	3.85%	3.61%	8.98%	7.85%	6.74%
actor-direction	1.52%	1.24%	4.67%	1.15%	0.87%	1.81%	0.90%	1.74%
actor-location	0.81%	0.56%	3.16%	1.05%	1.22%	0.65%	1.83%	1.32%
verb-object	8.21%	25.32%	9.59%	18.34%	22.04%	9.60%	11.78%	14.98%
verb-direction	1.11%	2.08%	0.18%	0.59%	3.74%	0.73%	0.00%	1.20%
object-direction	13.39%	11.72%	9.78%	8.85%	3.77%	8.65%	2.84%	8.43%
object-location	1.47%	7.15%	5.61%	5.14%	3.02%	1.47%	7.52%	4.48%
actor-verb-object	0.28%	2.14%	1.25%	1.16%	1.47%	1.05%	1.45%	1.26%
actor-verb-direction	0.02%	0.15%	0.00%	0.01%	0.22%	0.04%	0.00%	0.06%
actor-object-direction	0.60%	0.57%	1.79%	0.27%	0.18%	0.89%	0.25%	0.65%
verb-object-direction	0.14%	0.96%	0.08%	0.20%	1.76%	0.27%	0.00%	0.49%
mean	2.67%	5.20%	4.51%	3.67%	3.86%	3.06%	3.09%	3.72%

$$\frac{\left|\left(\bigcup_i \text{independent}_i\right) \cap \text{joint}\right|}{|\text{joint}|}$$

analysis	1	2	3	4	5	6	7	mean
actor-verb	73.84%	84.17%	29.37%	69.92%	84.69%	62.24%	67.68%	67.42%
actor-object	58.13%	70.01%	12.69%	48.66%	58.19%	50.56%	42.26%	48.64%
actor-direction	70.57%	76.98%	44.23%	53.25%	92.59%	81.29%	60.79%	68.53%
actor-location	77.87%	78.21%	43.04%	60.20%	85.62%	84.59%	56.04%	69.37%
verb-object	87.64%	91.44%	47.81%	79.92%	95.24%	72.63%	82.00%	79.53%
verb-direction	78.78%	91.95%	38.70%	57.35%	91.39%	76.85%	86.66%	74.53%
object-direction	70.03%	38.04%	71.44%	73.35%	63.40%	81.09%	64.41%	65.97%
object-location	72.66%	97.45%	95.52%	87.33%	75.24%	68.75%	57.12%	79.15%
actor-verb-object	19.80%	21.10%	9.13%	12.40%	52.40%	42.90%	13.60%	24.48%
actor-verb-direction	67.37%	69.78%	28.90%	44.15%	95.21%	56.72%	53.88%	59.43%
actor-object-direction	57.39%	21.04%	10.21%	38.26%	52.20%	43.39%	41.94%	37.78%
verb-object-direction	57.12%	66.48%	47.87%	42.93%	82.05%	43.58%	45.87%	55.13%
mean	65.93%	67.22%	39.91%	55.64%	77.35%	63.72%	56.02%	60.83%

Figure S28: Per-subject quantitative comparison of the brain regions indicated by searchlight of the independent classifiers to the joint classifiers, for all constituent pairs and triples, together with means across subject, means across analysis, and means across both. (top) The percentage of voxels in the union of the constituents for the independent classifier also in the intersection. (bottom) The percentage of voxels in the joint classifier that are shared with the independent classifier.

$$\frac{\left|\bigcap_i \text{independent}_i\right|}{\left|\bigcup_i \text{independent}_i\right|}$$

analysis	1	2	3	4	5	6	7	mean
actor-verb	1.21%	6.10%	2.19%	2.66%	1.88%	1.88%	3.95%	2.84%
actor-object	1.11%	2.61%	1.67%	3.84%	4.32%	1.98%	2.24%	2.54%
actor-direction	0.50%	0.70%	1.78%	0.55%	0.50%	3.14%	0.95%	1.16%
actor-location	0.80%	1.31%	2.51%	1.93%	2.04%	3.09%	2.77%	2.06%
verb-object	5.54%	5.26%	4.82%	6.72%	11.04%	5.09%	3.89%	6.05%
verb-direction	3.95%	3.25%	2.35%	3.68%	4.27%	2.82%	4.98%	3.61%
object-direction	7.87%	1.93%	2.66%	3.14%	3.78%	5.42%	1.06%	3.70%
object-location	0.90%	1.72%	3.03%	3.09%	3.95%	2.14%	1.67%	2.36%
actor-verb-object	0.00%	0.76%	0.10%	0.75%	0.88%	0.21%	0.28%	0.42%
actor-verb-direction	0.00%	0.03%	0.00%	0.03%	0.00%	0.03%	0.00%	0.01%
actor-object-direction	0.00%	0.00%	0.03%	0.03%	0.00%	0.00%	0.00%	0.00%
verb-object-direction	0.29%	0.21%	0.17%	0.10%	0.15%	0.36%	0.10%	0.20%
mean	1.85%	1.99%	1.78%	2.21%	2.73%	2.18%	1.82%	2.08%

$$\frac{\left|\left(\bigcup_i \text{independent}_i\right) \cap \text{joint}\right|}{|\text{joint}|}$$

analysis	1	2	3	4	5	6	7	mean
actor-verb	58.59%	71.79%	57.89%	58.69%	57.79%	57.19%	50.00%	58.85%
actor-object	47.89%	55.70%	52.20%	52.90%	47.39%	55.70%	46.80%	51.22%
actor-direction	42.89%	39.80%	39.00%	34.30%	43.10%	49.89%	48.00%	42.42%
actor-location	25.00%	28.59%	21.39%	22.50%	42.89%	28.10%	25.50%	27.71%
verb-object	64.70%	67.90%	65.60%	68.10%	70.59%	69.09%	56.39%	66.05%
verb-direction	67.50%	47.69%	55.00%	62.00%	72.09%	64.79%	67.60%	62.38%
object-direction	54.40%	37.10%	58.69%	51.80%	56.10%	57.39%	53.40%	52.70%
object-location	45.10%	30.19%	37.10%	43.29%	42.29%	44.00%	29.69%	38.81%
actor-verb-object	61.79%	68.30%	54.40%	62.70%	62.50%	60.39%	54.70%	60.68%
actor-verb-direction	68.89%	52.60%	51.60%	49.70%	55.10%	61.70%	56.00%	56.51%
actor-object-direction	45.10%	37.39%	45.89%	27.00%	32.70%	42.29%	38.10%	38.35%
verb-object-direction	69.59%	38.29%	53.70%	58.59%	64.50%	60.69%	62.39%	58.25%
mean	54.29%	47.95%	49.37%	49.30%	53.92%	54.27%	49.04%	51.16%

Figure S29: Per-subject quantitative comparison of the brain regions indicated by the thresholded SVM coefficients of the independent classifiers to the joint classifiers, for all constituent pairs and triples, together with means across subject, means across analysis, and means across both. (top) The percentage of voxels in the union of the constituents for the independent classifier also in the intersection. (bottom) The percentage of voxels in the joint classifier that are shared with the independent classifier.

analysis	(a)	(b)	(c)	(d)	(e)	(f)	(g)
actor	4	126	504	18	72	576	4032
verb	3	168	504	24	72	576	4032
object	3	168	504	24	72	576	4032
direction	2	168	336	24	48	384	2688
location	2	84	168	12	24	192	1344
actor-verb	12	42	504	6	72	576	4032
actor-object	12	42	504	6	72	576	4032
actor-direction	8	42	336	6	48	384	2688
actor-location	8	21	168	3	24	192	1344
verb-object	9	56	504	8	72	576	4032
verb-direction	4	84	336	12	48	384	2688
object-direction	6	56	336	8	48	384	2688
object-location	6	28	168	4	24	192	1344
actor-verb-object	36	14	504	2	72	576	4032
actor-verb-direction	16	21	336	3	48	384	2688
actor-object-direction	24	14	336	2	48	384	2688
verb-object-direction	12	28	336	4	48	384	2688
sentence	72	7	504	1	72	576	4032

Figure S30: The single constituent, joint constituent pair, joint constituent triple, and independent sentence analyses are separated by horizontal lines. The number of classes and number of test samples for independent and joint analyses for corresponding constituent pairs and triples are the same. No classifiers were trained for the independent constituent pair and triple analyses as these used the single-constituent classifiers. The number of training samples for the sentence analysis is the hypothetical number for a joint classifier that was not trained; only independent classification was attempted due to insufficient training-set size. (a) Number of classes. (b) Number of training samples per subject, fold, and class. (c) Number of training samples per subject and fold (columns a times b or column e times 7). (d) Number of test samples per subject, fold, and class. (e) Number of test samples per subject and fold (columns a times d). (f) Number of test samples per subject (column e times 8). (g) Number of test samples (column f times 7).