A Appendix

A.1 Audio transcription and alignment

The audio track of each movie was first annotated by commercial services (`Rev.com` and `HappyScribe.com` depending on the movie) and manually corrected by trained annotators. A custom tool was developed to refine the alignment via an auditory spectrogram of 4 seconds at a time and slowed-down audio track. Annotators were instructed to adjust the onset and offset of every word to align with the spectrogram and their perception of when the word started and ended. The audio annotation tool automatically played the audio segment corresponding to each word to allow annotators to verify their work. As the audio was played a line marked the location of the audio sample in the spectrogram in real time.

Since speech recognizers often misused or missed critical punctuation marks, these were inserted by annotators manually. Sentences were then manually segmented. Annotators were instructed not to use abbreviations, even if they are common. Annotators marked audio segments that consisted of overlapping speech or signing. These were removed from the dataset. All foreign language was marked and removed from the dataset. Annotators were instructed to transcribe literally, i.e, contractions were used in the transcript only when spoken as such. Similarly, foreshortened words, e.g., goin' vs going, were transcribed as such when used by speakers. Cardinal numbers were spelled out. Longer numbers were spelled out as spoken, including conjunctions such as "and". All overheard words were transcribed, even when they could not easily be localized on the spectrogram, for example, short words such as "to" can sometimes be heard but no specific segment of the spectrogram seems to correspond uniquely to such words. In this case annotators were asked to mark their onset and offset as they heard the words. Transcripts are as spoken, without correction, even when the speaker erred omitting a word or using a word inappropriately.

A.2 Feature annotation

We extract 16 features that were included in the analyses (see Extended Figures table 4).

Visual features The visual scene scalar features were extracted from the middle frame presented during a word utterance via OpenCV 4.4.0 [50]. Brightness was quantified as the average pixel HSV value channel. Flow vectors were computed as dense optical flow over grey-scale frames via the OpenCV `calcOpticalFlowFarneback` function (pyramid scale 0.5, 5 levels, window size 11, 5 iterations, pixel neighborhood of 5, and smoothing of 1.1). Number of faces per-frame was estimated via the OpenCV `CascadeClassifier` function with the Haar cascade frontal face default classifiers over gray-scale frames (scale factor: 1.1, minimum neighbours: 4). The first frame of every word utterance was mean-normalized and than passed through a pretrained ResNet-50 object detector (Torchvision 0.6.1) to compute a visual vector image embedding (size 2,048) as the last feature layer of the model.

Auditory features The auditory scalar features were collected with the Python Librosa package (0.7.2) [51], an open source audio analysis library. Sound intensity and mean frequency of the audio track during word utterance were estimated, as well as their change relatively to the preceding 500*ms* window. The average intensity of the audio segment was computed as the root-mean-square (RMS) (`rms` function, frame and hop lengths 2048 and 512 respectively) of that segment. Pitch was extracted using Librosa's `piptrack` function over a Mel-spectrogram (sampling rate 48,000 Hz, FFT window length of 2048, hop length of 512, and 128 mel filters). Auditory vector embeddings were computed as the flattened log-Mel-spectrogram of the 500*ms* word utterance window (size $128 \times 47 = 6016$).

Language features We used a state-of-the-art syntactic parser, Stanford NLP Group's Stanza [52], to parse every sentence. POS tags were recorded for every word. Surprisal was quantified as the negative-log word probability. Word probabilities were estimated by a transformer model. GPT-2 probabilities were computed via GPT-2 large using the Hugging Face Transformers 3.0.0 library [53]. Word particle surprisal were combined by summation.

All Universal Dependency features were inferred using the standard English model of the Stanza Natural Language Processing toolkit [52] and then manually corrected via a single trained annotator over the course of a year.

Speaker annotation Annotators doing speaker identification were instructed to use the characters' full names, insofar as they are known. If a character is unnamed, the annotator may identify them with a brief description of their role.

Occasionally, a character had another identity that they went by. In Spider-Man: Homecoming, the AI in Peter's suit is known for more than half the movie as "suit lady," until Peter finally decides to give her the name "Karen." In such situations, the annotator marked both identities, with whichever identity they decide is primary listed first, and the secondary identity in parentheses. So, in the above example, Peter's AI is annotated as "Karen (suit lady)"

Because of our data set, we deal with quite a lot of super heroes with secret identities. If a super hero was in costume, annotators identified them by their super hero name. Out of costume, they were identified by their birth name. When they are partially in costume (say, they're in costume, but they've taken off their mask), annotators identified them by their super hero name, followed by their birth name, separated by a forward slash: e.g. Spider-Man / Peter Parker

In situations where one character is pretending to be another, the guidelines bear some resemblance to the guidelines for heroes that are partially in costume. Annotators identified them by the person being imitated, followed by the true identity of the character, separated by a percent symbol. So, for a good part of the movie Megamind, the titular character is pretending to be a museum curator named Bernard. Dialog spoken by him during these moments should be annotated as "Bernard % Megamind."

Lines that had problems and therefore that need special attention can be identified using an asterisk. Two of the most common situations where this cropped up were when multiple characters were speaking in unison, or when a "sentence" actually contains utterances from multiple characters. In the former situation, these were identified with the line with `* multiple speakers`. In the latter situation, both speakers were annotated, with an asterisk between them e.g. "Peter Parker * Tony Stark," and an asterisk was added to the line of dialog at the point where one of them stops speaking and the other begins.

A.3 Task and stimuli

Movies were extracted from DVDs and are unchanged other than being re-encoded to a fixed frame rate (23.976 fps). Transcripts, and all annotations described in this work will be made publicly available. Due to copyrights prohibiting the release of the raw stimuli (movies) source material, multiple audio-visual sample clips and tools allowing users to verify alignment of their own movie copies will be publicly provided.

Movies were shown in full to each subject. Movies were displayed via a custom video player created in Matlab 2018b. The player ensured that the presentation was at a fixed frame rate to keep the audio and video synchronized. The presentation of movies was accompanied by regular electrical triggers sent to the neural recording system to enable accurate temporal alignment between the movie and the neural data. A 15.4 inch (resolution 2880×1800) Apple MacBook Pro Retina was placed 60-100cm in front of the subject. Subjects adjusted the volume and paused/resumed the movie as needed. The movie was paused by the experimenter any time someone entered the room or when subjects were distracted and was resumed when subjects could direct their full attention back to the movie. Subjects could freely change position, but were instructed by the experimenter, who watched the movies with the subjects, to remain focused on the stimulus or pause the movie. Subjects did not speak during the presentation of the movie nor did they overhear any other speech other than that found in the movie.

A.4 Data acquisition and signal processing

Clinicians implanted subjects with intracranial stereo-electroencephalographic (SEEG) depth probes containing 6-16 0.8 mm diameter 2 mm long contact electrodes (Ad-Tech, Racine, WI, USA) recording Intracranial Field Potentials (IFPs) with 1.5 mm separation. Each subject had multiple (12 to 18) such probes implanted in locations determined by clinical concerns entirely unrelated to the experiment. Data was recorded using XLTEK (Oakville, ON, Canada) and BioLogic (Knoxville, TN, USA) hardware with a sampling rate of 2048 Hz.

During movie presentation, triggers were sent to a separate channel on the neural recording device via a USB connection to a dedicated trigger box (Measurement Computing USB-1208FS) using the Psychtoolbox 3 Matlab package. Each pulse was logged with both its wall-lock timestamp and its movie timestamp. Individual triggers were sent every $100ms$. Specific events (movie start, pause, resume, and end) were marked by bursts of triggers (10, 8, 9, and 11 respectively) separated by 15ms. All triggers consisted of a 15ms electrical burst at a magnitude of 80mV. An automated tool found triggers and aligned the movie and neural data.

A.5 Cortical surface extraction and electrode visualization

For each subject, pre-operative T1 MRI scans without contrast were processed with FreeSurfer's `recon-all` function with `-localGI`, which performed skull stripping, white matter segmentation, surface generation, and cortical parcellation [54–73]. iELVis [74] was used to co-register a post-operative fluoroscopy scan to the preoperative MRI. Electrodes were manually identified using BioImageSuite [75], and then assigned to one of 68 regions (according to the Desikan-Killiany atlas [46]) using FreeSurfer's automatic parcellation. The alignment to the atlas was manually verified for each subject. One subject had a large frontal lesion in the right hemisphere that prevented alignment to an atlas. Electrodes from this subject were included in all analyses except for region analyses and they were not plotted on the brain.

Corrupted signal electrodes ($n = 114$) with extensive durations of static signal recordings were manually removed from consideration prior to any downstream analysis. For depth electrodes in the white matter, if they were within 1.5 mm of the gray-white matter boundary, they were projected to the nearest point on that boundary, and were labeled as coming from that region (for the purposes of region significance analyses). Of the 1,688 total electrodes, 1,414 of the electrodes were able to placed in this way into a particular region. The relevant region analyses are shown in fig. 2h-i, fig. 3f-h, fig. 12e-f, fig. 4b, fig. 2b, fig. 3b, fig. 5e.

This procedure is very similar to the post brain-shift correction methods used for electrocorticography electrodes [76]. For solely visualization purposes, all electrodes identified to lie in the gray matter or on the gray-white matter boundary were first projected to the pial surface (using nearest neighbors), and then mapped to an average brain (using Freesurfer's fsaverage atlas) for the visualizations shown in the main text.

A.6 Word responsiveness

To determine the word responsiveness of an electrode, we compared pre-onset windows to post-onset windows (fig. 10). Precisely, we compared the mean activity in a 100ms window before word onset to the activity in a 100ms window after word onset with a two-tailed paired t-test. The windows were separated by an interval of 1s. This test was performed for absolute offsets of $[-0.5s, -0.4s, -.3s, -.2s, -.1s]$ (fig. 10). This is done to account for the fact that any one offset may "miss" the neural response by chance. An electrode is *word responsive* if at least one of the tests shows a significant (after correction for multiple comparisons) difference between pre- and post-onset activity. In such cases, we report the significance of the t-test with the lowest p-value.

A.7 Testing difference between conditions

When determining the significance of the difference between two conditions (fig. 2c, fig. 3b, fig. 12a,c,d), we used a two-tailed t-test to compare the mean activity in a 100ms window for the two conditions. Five t-tests are performed, at absolute offsets of $[0s, 0.1s, 0.2s, 0.3s, 0.4s]$ and we say that the two conditions result in different neural responses if there exists a test for which there is a significant difference, after correction for multiple comparisons. In such cases, we report the significance of the tests with the lowest p-value. As in the above section, this is done to account for the fact that any one of the tests may miss the difference between the two conditions by chance.

A.8 Linear decoding

Model The model is a logistic regression.

Data pre-processing Neural is decimated by a factor of 10. Data is normalized to 0 mean and unit standard deviation. Normalization is done such that no data-leakage occurs (see below).

Dataset The sentence-onset decoding task requires the model to distinguish between neural activity from an interval in the movie during which a sentence is beginning versus an interval during which no speech is occurring. To obtain positive examples, for every sentence onset, we extract 2s of neural activity, centered on the sentence onset. To obtain negative examples, we divide the movies into 3s segments, and filter for segments that do not overlap with any speech time-stamps. The size of 3s guarantees that there is at least a 500ms buffer between every positive example and every negative example (see below). The dataset is balanced so that an equal number of negative and positive examples occur. Data is drawn from all recorded movies per subject.

Training We are interested in answering the question, how does decodability vary across time? To this end, we divide each example into 250ms intervals. Per each time interval, per electrode, we train our model. Training was done on a single NVIDIA Titan RTXs (24GB GPU Ram) with 80 CPU cores.

Evaluation Per electrode, we create an 80/20 train/test split. The model performance is reported on the test set. Train/test splits are shared between electrodes in the same subject. In fig. 4b,d, and e, we select the top 10 electrodes with the highest score on the train-set (5-fold cross-validation) per region, and report the performance of these electrodes on the test set. The same is done in fig. 2b,d-e and fig. 3b,d-e.

A.9 Part of speech modulates activity

Parts of speech are of particular importance for their fundamental role in linguistics and natural language processing (NLP). Indeed, the two word classes, nouns and verbs, are widely recognized to be among the few linguistic universals [77, 78]. Part-of-speech was a significant predictor in the example electrodes shown in fig. 2 and fig. 3. Given their importance in language, we directly compared the responses to nouns versus verbs (fig. 12). fig. 12a shows the responses of an example electrode located in the left superior temporal gyrus (inset) which showed stronger responses to verbs compared to nouns.

The GLM analysis showed that there were no electrodes which exhibited activity exclusively modulated by part-of-speech. Instead, the neural activity was captured by multiple features as shown in the previous examples. fig. 12b shows that the main feature for this electrode is the index in sentence, followed by the part-of-speech and volume. Indeed, after separating the responses according to the position in the sentence, there was a small but significant difference between nouns and verbs for sentence midsets and offsets but not for sentence offsets (fig. 12c). The differences between nouns and verbs persisted across high and low volumes (fig. 12d). There were no electrodes for which a difference in part-of-speech was observed across all sub-samplings for all features. But there were 83 electrodes for which part of speech has a significant ($p < 0.05$, Bonferroni corrected) beta coefficient in the GLM analysis. fig. 12e shows the exact location of these electrodes and fig. 12f shows the fraction, per region, of the part of speech significant electrodes. We also found that the noun-verb distinction is linearly decodable fig. 3, with significant decoding performance distributed across the brain fig. 3a, and with the highest decoding performance observed in the frontal lobe and cingulate (fig. 3b-e).

Finally, we observed a difference in the magnitude and timing of the peak neural response between nouns and verbs (fig. 13). For each electrode, we computed the mean of the neural response, averaged across all words. Restricting our attention to those electrodes which show at least a moderate neural response (Cohen's $d > 0.1$, see section 3), we can compute the peak of that mean response (fig. 13b) and observe that it is lower in the case of verbs at sentence onsets. ($\mu \approx 32.6, \sigma = 27.7\ \mu V$ for verbs, $\mu \approx 35.5, \sigma = 29.7\ \mu V$ for nouns), but higher in the case of verb midsets ($\mu = 34.1, \sigma = 25.9\ \mu V$ for verbs, $\mu = 30.5, \sigma = 26.4\ \mu V$ for nouns). We also find the timing (fig. 13c) of the sentence midset peaks and observe that it is earlier in the case of verbs ($\mu \approx 293, \sigma = 255\ ms$ for verbs, $\mu \approx 426, \sigma = 315\ ms$ for nouns).

Subj.	Age	Sex	Movies	Time (h)	# Sentences	# Words	# Lemmas	# Electrodes	# Probes
1	19	M	7, 18, 19	6.14	4054	29468	5908	154	13
2	12	M	2, 3, 4, 8, 9, 17, 21	15.49	9092	60958	12243	162	47
3	18	F	5, 11, 12	9.50	4845	32959	6156	134	12
4	12	F	10, 13, 15	5.06	3758	25394	5300	188	15
5	6	M	7	1.45	1162	8457	1892	156	12
6	9	F	6, 13, 20	8.02	3524	21455	4544	164	12
7	11	F	5, 13	3.36	3152	20237	3808	246	18
8	4	M	14	0.96	718	4218	804	162	13
9	16	F	1	1.95	1412	9846	1956	106	12
10	12	M	5, 16	3.93	3506	23408	4048	216	17

Table 2: All subjects language, electrodes and personal statistics. Columns from left to right are the subject's ID and information (age and gender), the the IDs of the movies they watched (corresponding to Extended Figures table 3), the cumulative movie time (hours), number of sentences, number of words (tokens) and number of unique lemmas (canonical word forms), as well as the number of probes the subject had and their corresponding number of electrodes..

B Supplementary figures

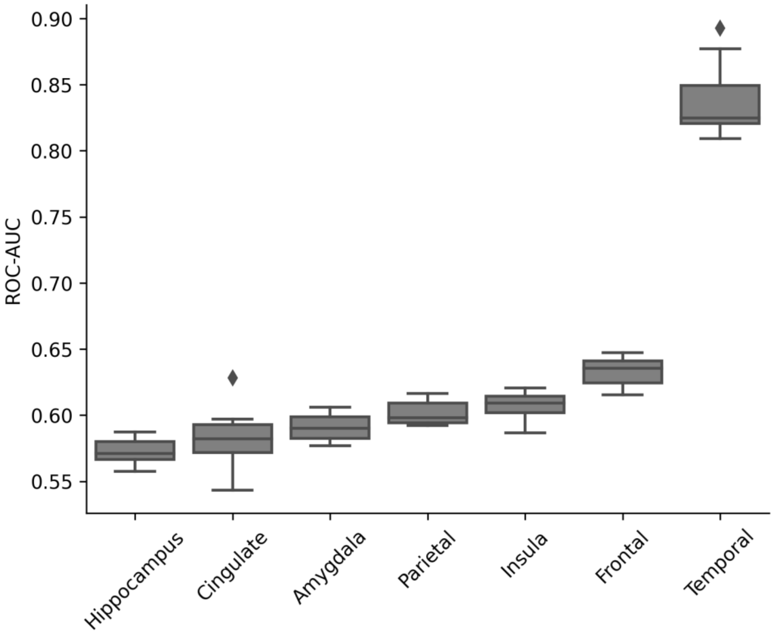

Figure 1: **Decodability of sentence onsets per region.** After decoding sentence onsets per electrodes (see fig. 4), we find distribution of the peak *test* ROC-AUC scores in each region, for the 10 electrodes in each region with the highest cross-validation ($k_{\text{folds}} = 5$) ROC-AUC on the *train* set. Boxes show quartiles and whiskers show 1.5× the interquartile range. Outliers shown as points beyond the whiskers.

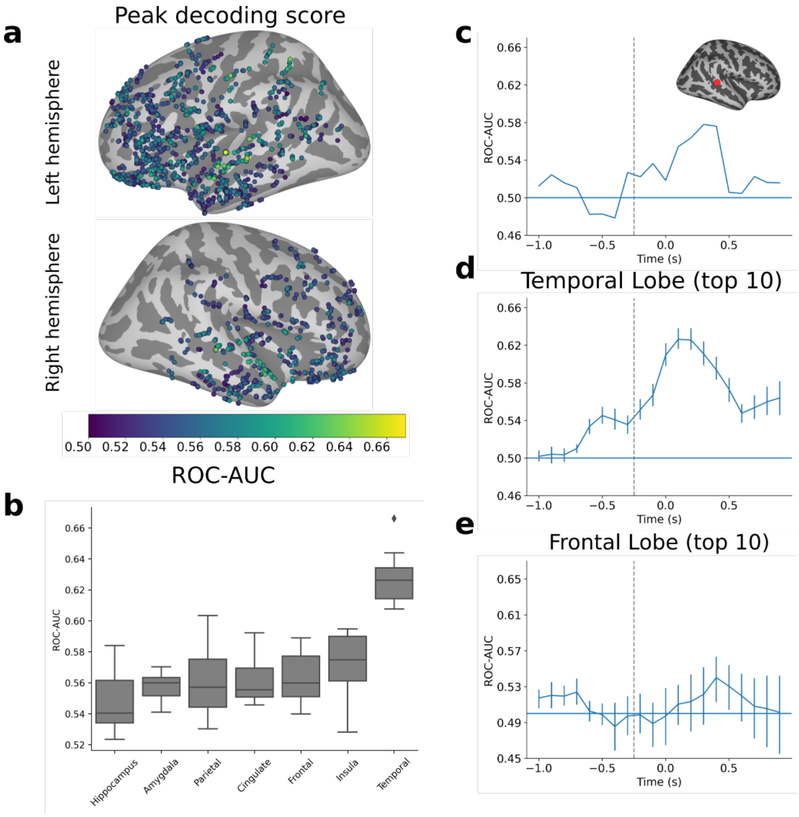

Figure 2: **Word onsets are linearly decodable and reveal the time course of language processes in the brain.** We perform the same analysis as shown in fig. 4, but for word-onsets, instead of sentence-onsets only. A linear decoder is trained to classify portions of the movies according to whether or not speech is occurring, based on the corresponding neural activity. This decoding is done for activity in a 0.25s window, which shifts in 0.1s increments from -1s before word-onset to 1s after word-onset. The spatial distribution of decoding scores, shown in (**a**) and (**b**), after a max has been taken over all windows, shows that word onsets are most decodable in the temporal and frontal lobes. Decodability, as a function of time, shown in (**c**), (**d**), and (**e**), reveal that some word onset information is processed before word onset enters the decoding window (dashed grey line). Averaging over time across the top 10 electrodes per region, as in (**d**) and (**e**), reveals the mirrored time course of language processing.

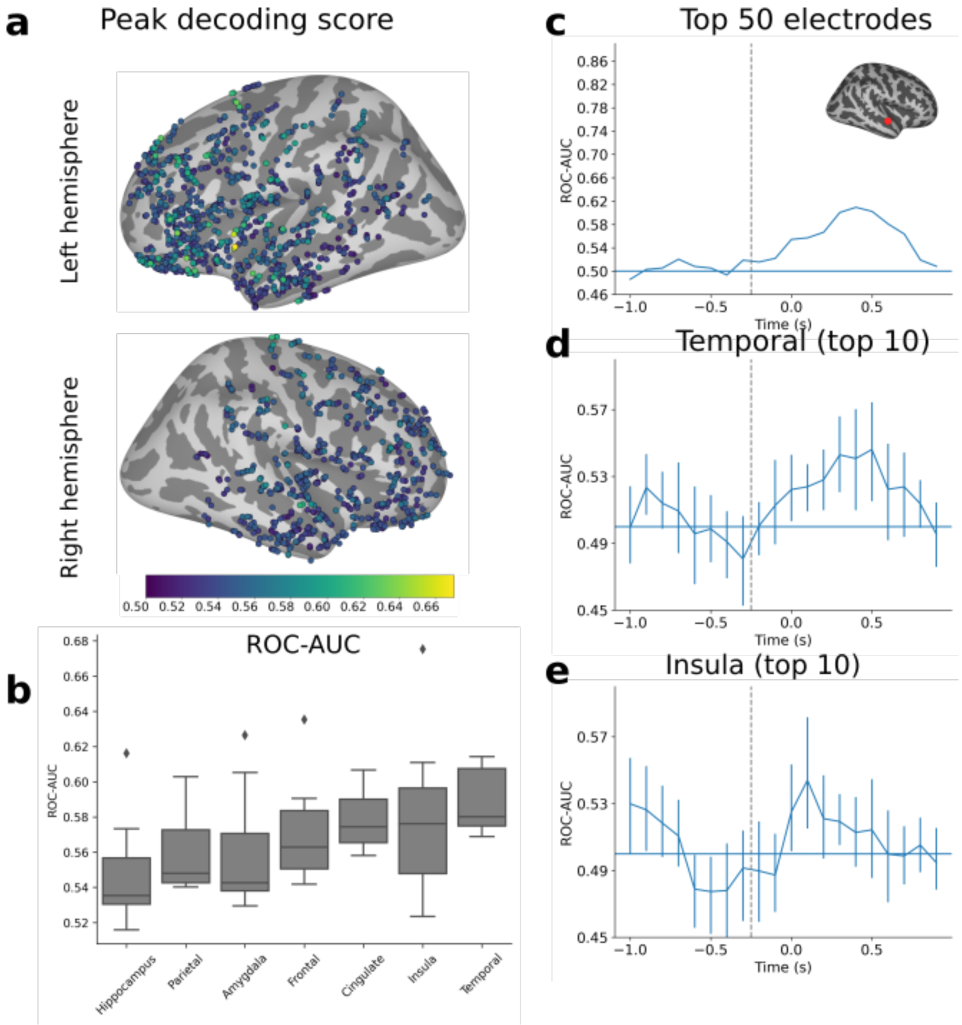

Figure 3: **Part of speech information is linearly decodable.** We perform the same analysis as shown in fig. 4 for nouns and verbs. A linear decoder is trained to classify words as either nouns or verbs, based on the corresponding neural activity. This decoding is done for activity in a 0.25s window in 0.1s increments. The spatial distribution of decoding scores, shown in (a) and (b), after a max has been taken over all windows, shows that part of speech is most decodable in the frontal, cingulate, insula, and temporal regions. Decodability, as a function of time, shown in (c, for an electrode in the superior temporal lobe), (d), and (e), reveal that some part of speech information is processed before word onset enters the decoding window (dashed grey line). Averaging over time across the top 10 electrodes per region, as in (c) and (d), reveals the time course of processing.

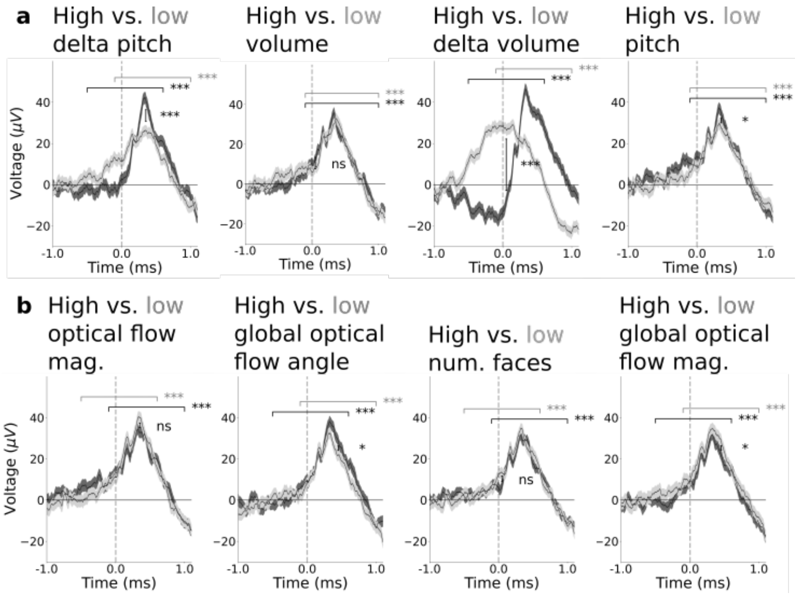

Figure 4: **Neural responses to word onsets are observable, even after controlling for visual and audio features. a**. Mean response to word onsets, after controlling for audio features for the same example electrode as shown in fig. 2. The same conventions as fig. 2c are followed. Vertical brackets and corresponding asterisks show the difference between conditions. Horizontal brackets and asterisks show the significance of the word onset response. **b.** Mean response to word onsets, after controlling for visual features. In both (a) and (b), significant response to word onset can be observed, even after controlling for audio and visual features respectively.

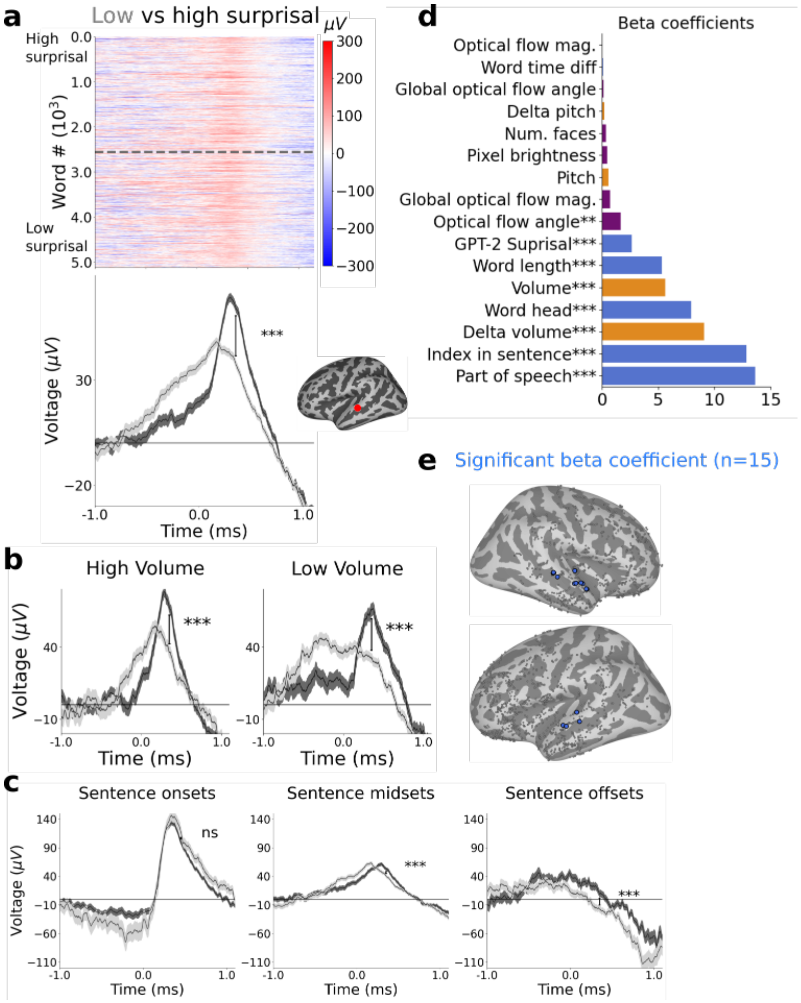

Figure 5: **Neural responses distinguish high and low surprisal. a.** Raster and mean plots aligned to word onsets for an example electrode in the right superior temporal gyrus (see inset in **d**; this is the same electrode as shown in fig. 12) separated by high and low surprisal. The difference between high and low surprisal words remains even after controlling for other features, such as volume (b) and position in sentence (c). GLM analysis reveals that activity in this electrode is modulated in part by surprisal, as well as by other features (d). There are 10 electrodes where part of speech has a significant beta-coefficient; these are all located in the superior temporal lobe (e).

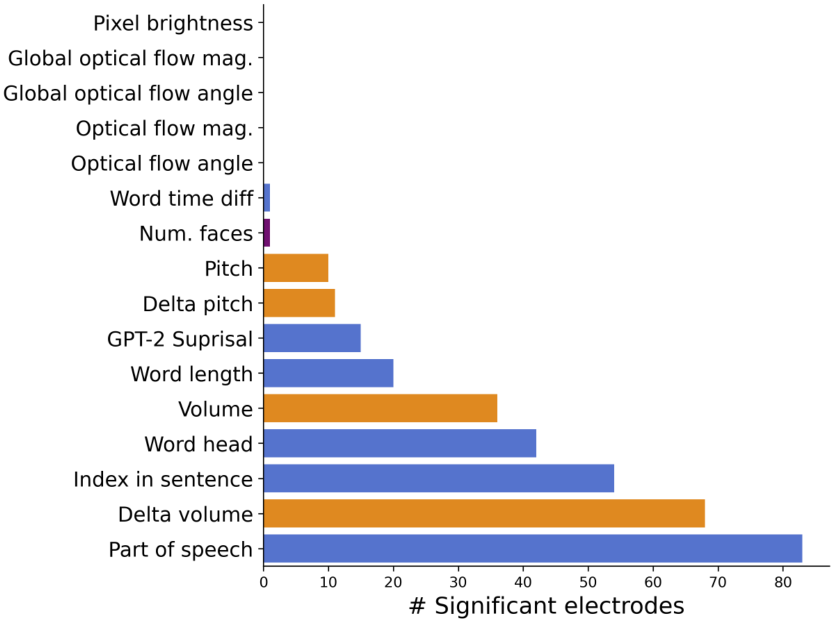

Figure 6: **The other factors which influence activity in part-of-speech-sensitive electrodes**. An electrode is said to be sensitive to part-of-speech, if a GLM fitted to mean neural activity has a significant beta coefficient ($p < 0.05$, after corrections for multiple comparisons) for the part-of-speech feature. Among all such part-of-speech sensitive electrodes ($n = 83$) , the number of electrodes that have other significant beta coefficients is shown.

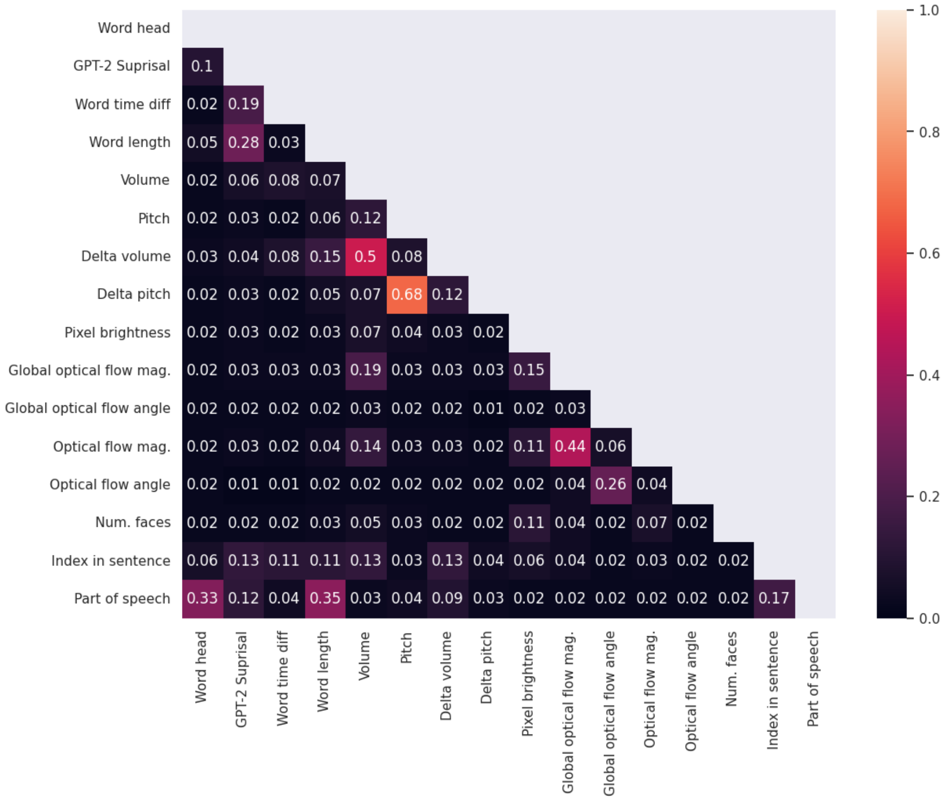

Figure 7: The (absolute value) of Pearson's r between input features, averaged across movies.

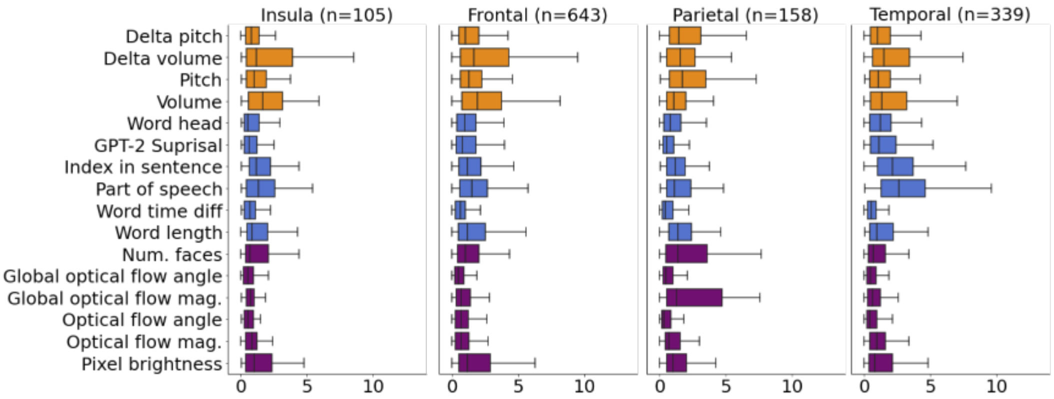

Figure 8: Magnitude of beta coefficients, averaged per region.

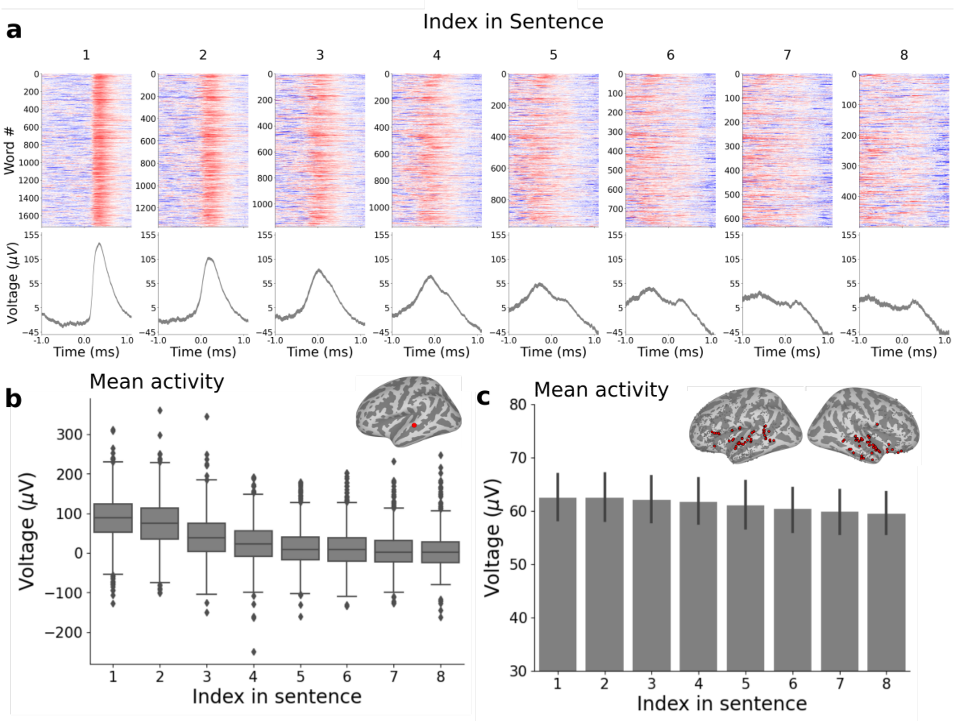

Figure 9: **Neural response decreases as a function of position in the sentence**. Making a more fine-grained examination of sentence position, we observed a trend in which mean activity decreased monotonically with the index in the sentence. (**a**) The neural response per index in sentence is shown for the first eight sentence positions for an electrode in the left temporal lobe (same electrode as shown in fig. 12). (**b**) The mean activity for this same electrode (location shown in inset) is taken for a [0ms,500ms] window after word onset. The box shows the quartiles, while the whiskers show 1.5 × the interquartile range, over all words at a given position. (**c**) Taking the mean of the magnitude over this same window for all word responsive electrodes shows the same trend. Error bars show a 95% confidence interval over electrodes. A word-responsive electrode is defined, as in fig. 2, as an electrode that shows a significant difference between pre- and post- onset activity. Here we restrict our attention to those electrodes ($n = 111$, locations shown in inset) for which this difference has at least a moderate effect size (Cohen's $d > 0.1$). Note that we do not believe this result stands in opposition to previous findings, such as in [79], foremost because we consider a much different distribution of sentences in our work. The sentences shown to subjects in this work cover a wide variety of forms, and importantly, are usually part of a longer dialogue. To make a direct comparison with previous studies of sentence processing, a more fine-grained inventory of sentence types should be made over the movie transcripts.

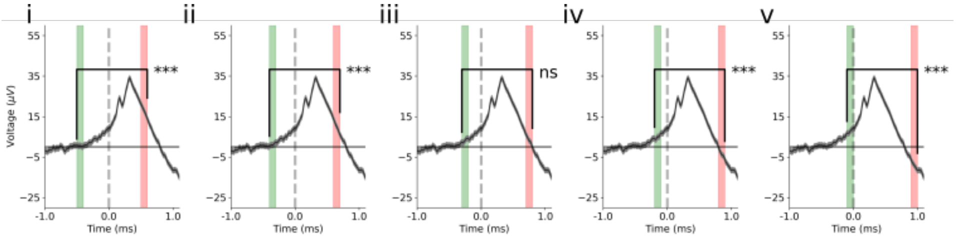

Figure 10: **Schematic of word-responsiveness testing procedure**. We test for word responsiveness at five different points (i-v). The grey line shows mean neural response, averaged across a movie. Shading shows standard error. At each point, a two-tailed paired t-test is performed between the mean activity in a pre-onset (green) and a post-onset (red) window of 100ms. We use multiple tests to account for the fact that sometimes the difference in activity may be 0 simply due to the absolute offset of the windows (this is the case for iii). We say that an electrode is word-responsive, if there is at least one test for which there is a significant difference between pre- and post- onset activity, after correcting for multiple comparisons.

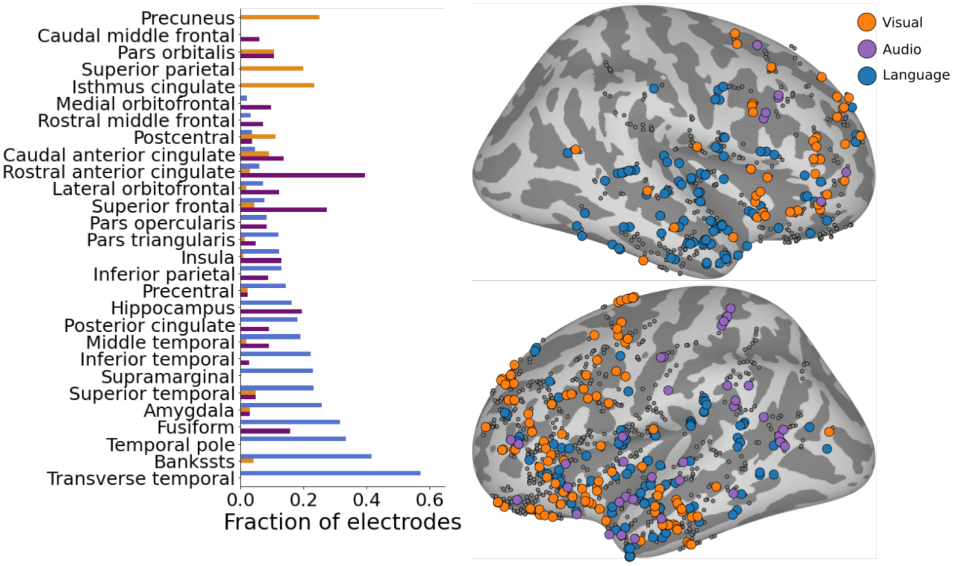

Figure 11: **Unimodal responsive electrodes**. We categorize features as either *visual*, *audio*, or *language*. For each electrode, we use the GLM analysis to determine whether a given electrode's activity has a significant (after Bonferroni correction) response for features from a single category, to the exclusion of the other categories.

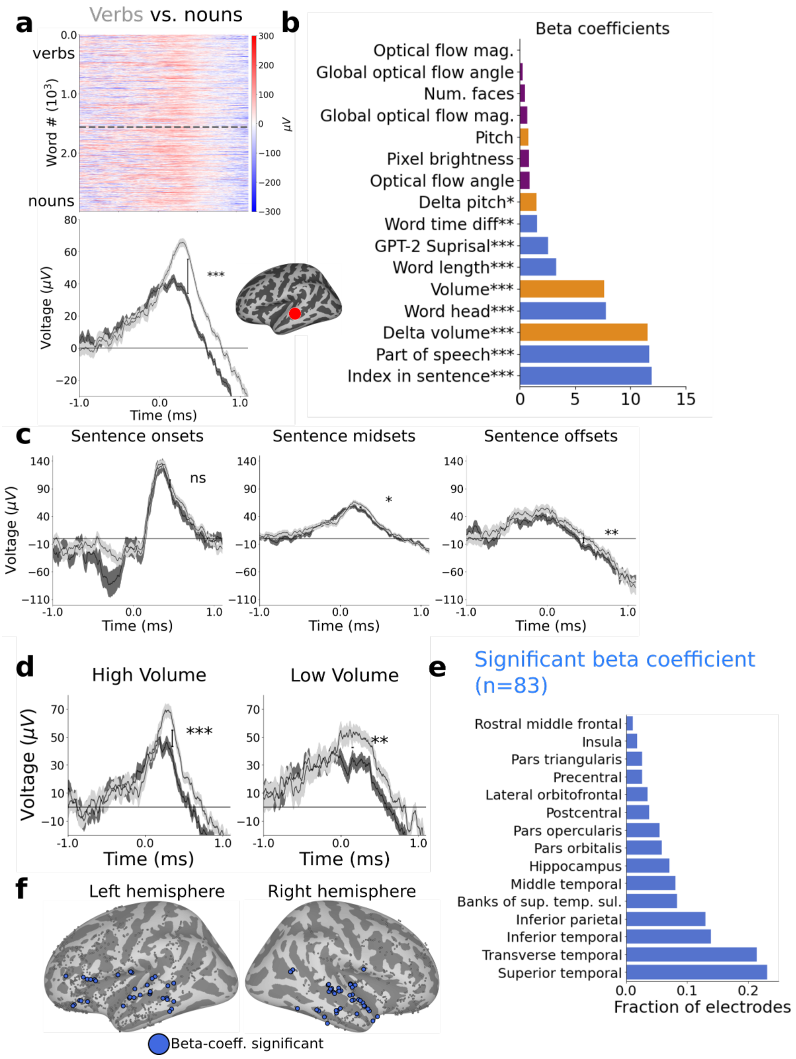

Figure 12: **Neural responses distinguish nouns and verbs. a.** Raster and mean plots aligned to word onsets for an example electrode in the left superior temporal gyrus (see inset) separated by nouns (bottom in raster plot, light grey in mean plot) and verbs (top in raster plot, dark grey in mean plot). **b.** GLM analysis reveals that activity in this electrode is modulated by part of speech, as well as by other features. **c.** For this electrode, a significant difference between nouns and verbs does not remain for the sentence onsets condition, after sub-sampling over sentence position. **d.** But, a difference does remain for all sub-sampled conditions, when controlling for other features, such as volume. Using the GLM analysis, allows us to judge the influence of part-of-speech on a per-word basis. **e.** The fraction of electrodes, per region, of electrodes where part of speech has a significant beta-coefficient (total $n = 83$); these are mainly located in the temporal and frontal lobes. **f.** The exact location of these electrodes (blue) projected onto the surface of the brain.

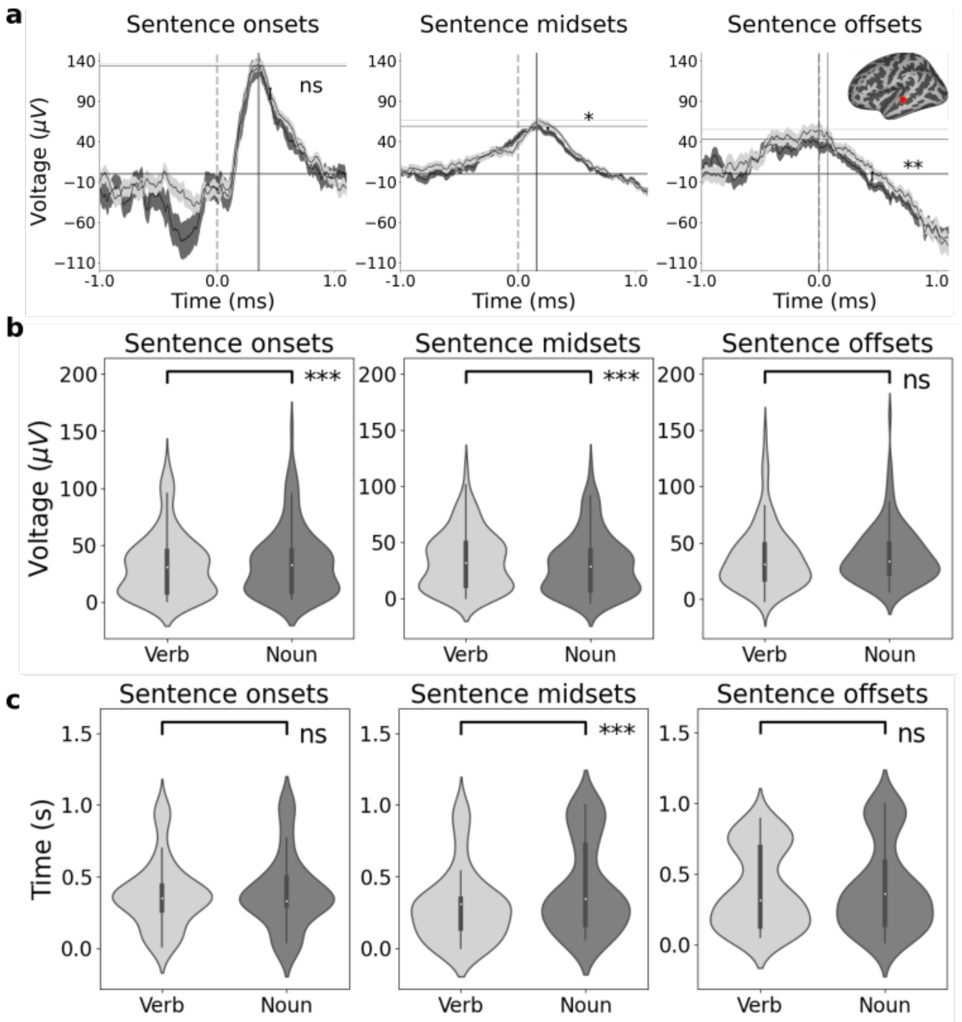

Figure 13: **Noun vs. verb peak amplitude and timing.**. For each electrode, we consider the mean signal. See, for example, (a) which shows the mean activity for an electrode in the STG (the same electrode shown in fig. 12). For an electrode, we find the amplitude (horizontal lines) of the peak mean activity and the timing of the peak (vertical lines). Across many electrodes, we observe a difference in the peak amplitudes such that nouns induce a higher response than verbs for sentence onsets, while verbs induce a higher response for offsets and midsets. The electrodes in (b) and (c) are those electrodes which respond to language (see fig. 2d), with the additional condition that the language response have at least moderate effect size (Cohen's d > 0.1).

						Unique		Unique		Unique
#	**Movie**	**Year**	**Time (s)**	**# Sentences**	**# Words**	**words**	**Nouns**	**nouns**	**Verbs**	**verbs**
1	Antman	2015	7027	1412	9846	1956	1370	712	1538	581
2	Aquaman	2018	8601	1003	7218	1563	1066	517	1094	508
3	Avengers: Infinity War	2018	8961	1372	8479	1780	1081	608	1294	485
4	Black Panther	2018	8073	1139	7571	1628	1084	544	1199	506
5	Cars 2	2011	6377	1801	11404	2060	1576	737	1649	563
6	Coraline	2009	6036	933	5428	1251	759	407	817	353
7	Fantastic Mr. Fox	2009	5205	1162	8457	1892	1240	690	1240	490
8	Guardians of the Galaxy 1	2014	7251	1104	8241	1799	1101	615	1235	521
9	Guardians of the Galaxy 2	2017	8146	1180	9332	1839	1210	623	1368	533
10	Incredibles	2003	6926	1408	9369	1966	1234	659	1545	582
11	Lord of the Rings 1	2001	13699	1424	10538	2011	1470	681	1480	595
12	Lord of the Rings 2	2002	14131	1620	11017	2085	1593	760	1587	631
13	Megamind	2010	5735	1351	8833	1748	1183	610	1340	496
14	Sesame Street Ep. 3990	2016	3440	718	4218	804	716	233	674	211
15	Shrek the Third	2007	5568	999	7192	1586	989	568	1072	418
16	Spiderman: Far From Home	2019	7764	1705	12004	1988	1442	660	1755	555
17	Spiderman: Homecoming	2017	8008	1993	12258	2107	1591	795	1794	569
18	The Martian	2015	9081	1421	11360	2210	1781	826	1686	630
19	Thor: Ragnarok	2017	7831	1471	9651	1806	1183	604	1440	546
20	Toy Story 1	1995	4863	1240	7194	1545	1039	561	1015	388
21	Venom	2018	6727	1301	7859	1527	892	509	1200	427

Table 3: Language statistics for all movies. Columns from left to right are the movie's ID, name, year of production, length (seconds), number of sentences, number of words (tokens), number of unique words (types), number of nouns, number of unique nouns, number of verbs and number of unique verbs.

#	Feature	Category	Description
1	Pixel brightness	Visual	The mean brightness computed as the average HSV value over all pixels
2	Global optical flow magnitude	Visual	A camera motion proxy. The maximal average dense optical flow vector magnitude
4	Optical flow magnitude	Visual	A large displacement proxy. The maximal optical flow vector magnitude
5	Optical flow angle	Visual	The orientation (degrees) of the above flow vector
6	Number of faces	Visual	The maximal number of faces per frame
7	Volume	Auditory	Average root mean squared watts of the audio
8	Mean pitch	Auditory	Average pitch of the audio
9	Delta volume	Auditory	The difference in average RMS of the 500ms windows pre and post word onset
10	Delta pitch	Auditory	The difference in average pitch of the 500ms windows pre and post word onset
11	GPT-2 surprisal	Language	Negative-log transformed GPT-2 word probability (given sentence preceding context)
12	Word time length	Language	Word length (ms)
13	Word time difference	Language	Difference between previous word offset and current word onset (ms)
14	Index in sentence	Language	The word index in its context sentence
15	Word head	Language	The relative position (left/right) of the word's dependency tree head
16	Part of speech tag	Language	The word Universal Part-of-Speech (UPOS) tag

Table 4: **Extracted visual, auditory, and language features used to model the neural responses**. All scalar type features were used as regressors in the GLM analysis and all scalar and vector features were used as test set balancing features in the multi-confounds CNN analysis. The difference between 2 and 4 is that 2 is the magnitude of the averaged optical flow vector, with the average being taken over all optical flow vectors on the screen, whereas 4 is the magnitude of the largest individual optical flow vector on the screen.

C Data documentation

The brain recordings and annotations and annotations are released at the subject level, and can be thought of as the raw source, from which derivative machine learning datasets may be created, and for this reason we do not include any croissant meta-data. An example of a dataset derivation could be: segmenting the audio track by word boundaries and then training a decoding model to map for neural recordings to word identity. Another example could involve segmenting the recording into uniform intervals and then training a decoding model to predict average color on screen. We release the recordings in their entirety to allow for this flexibility.

The website contains the following assets:

1. `quickstart.ipynb` A quickstart IPython notebook
2. `localization.zip` Spatial position of electrodes
3. `subject_timings.zip` Wall clock time of triggers used for synchronization with movie
4. `subject_metadata.zip` Movie metadata
5. `electrode_labels.zip` Semantic ID for electrodes
6. `speaker_annotations.zip` Speaker IDs for movie audio
7. `scene_annotations.zip` Scene cut annotations for movies
8. `transcripts.zip` Pre-computed features for movies
9. `trees.zip` Universal Dependency parse trees for movie dialogue
10. `sub_<sub_id>_trial<trial_id>.h5.zip` Neural recordings in HDF5 format

D Responsibility, License, Hosting Plan

Authors bear all responsibility in case of privacy violations. Authors release the data under a CC BY 4.0 license.

Data will be hosted on MIT CSAIL servers and will be accessible at the url `https://braintreebank.dev/`. Backups will be kept across multiple machines. Hardware will be maintained by the MIT CSAIL Infrastructure Group: `https://tig.csail.mit.edu/`.