Natural language processing models reveal neural dynamics of human conversation

Jing Cai[1*], Alex E. Hadjinicolaou[2], Angelique C. Paulk[2], Ziv M. Williams[1,3,4,5*] & Sydney S. Cash[2,3,5*]

[1] Department of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA.
[2] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA.
[3] Harvard-MIT Division of Health Sciences and Technology, Boston, MA.
[4] Harvard Medical School, Program in Neuroscience, Boston, MA.

[5] These authors contributed equally.

[*] To whom correspondence should be addressed, e-mail:
jcai5@mgh.harvard.edu
scash@mgh.harvard.edu
zwilliams@mgh.harvard.edu.

Human verbal communication requires a rapid interplay between speech planning, production, and comprehension. These processes are subserved by local and long-range neural dynamics across widely distributed brain areas. How linguistic information is precisely represented during natural conversation or what shared neural processes are involved, however, remain largely unknown. Here we used intracranial neural recordings in participants engaged in free dialogue and employed deep learning natural language processing models to find a striking similarity not only between neural-to-artificial network activities but also between how linguistic information is encoded in brain during production and comprehension. Collectively, neural activity patterns that encoded linguistic information were closely aligned to those reflecting speaker-listener transitions and were reduced after word utterance or when no conversation was held. They were also observed across distinct mesoscopic areas and frequency bands during production and comprehension, suggesting that these signals reflected the hierarchically structured information being conveyed during dialogue. Together, these findings suggest that linguistic information is encoded in the brain through similar neural representations during both speaking and listening, and start to reveal the distributed neural dynamics subserving human communication.

Introduction

Our ability to communicate through language requires two distinct but complementary linguistic computations: comprehension and production. Speech comprehension involves a structured succession of processes that extract information from acoustic-phonetic signals allowing us to understand the meaning of sentences and to comprehend the thematic and contextual information being conveyed (*1-6*). In contrast, speech planning involves a reverse process whereby higher-order conceptual information is converted to motor planning features for articulation (*7-9*). Together, these processes are necessary for communicating information during conversation and often involve rapid alternations between speakers in every several seconds (*10, 11*).

Yet, how linguistic information is precisely represented in the brain during conversation or what common neural processes subserve speech production and comprehension have remained a challenge to understand, particularly during natural, free-flowing speech. Whereas certain brain areas, for example, have been shown to distinguish well-formed meaningful sentences from degraded or non-meaningful stimuli, suggesting their involvement in language processing (*12-15*), the neural process by which linguistic information may be represented in the brain has been difficult to study directly. Further, while there is often broad overlap between areas involved in language production and comprehension (*16-19*), whether common neural processes are involved in representing linguistic information when speaking and listening during conversation remains poorly understood. Finally, although certain areas of the brain have been implicated in conversation or turn-taking (*20-23*), little is known about whether or how brain activity relates to information conveyed during dialogue; especially when considered with the temporal dynamics of natural speech since the methods for analyzing natural communication are still limited. Our primary mode for understanding human language, for example, has been through highly structured tasks under artificial environments (*20, 24, 25*). Yet, these task settings are, by their nature, limited in revealing neural underpinnings of complex cognitive activities such as free-flowing speech.

The recent advancement of natural language processing (NLP) models based on artificial deep neural networks has provided a new prospective platform by which to begin studying continuous, natural linguistic interactions. These models have been shown to display high-level performative interaction with human subjects in conversations (*26*), and can achieve state-of-the-art benchmarks in comprehension-based tasks and question-answering (*27-29*). Indeed, these models are capable of capturing specific word sequences and their composition within phrases and sentences through hierarchical layers in a way that can be potentially compared to neural activity (*30, 31*). NLP models have also demonstrated high performance in explaining brain activity during passive listening (*32-34*), suggesting their neurobiological plausibility. Here, we utilize these models as artificial, hierarchically structured vectorized representations of language during natural dialogue in order to start investigating how linguistic information is represented in the brain during conversation.

Results

Neural recordings during natural conversation

Local field potential (LFP) recordings were obtained using semi-chronically implanted depth electrodes in six participants undergoing epilepsy monitoring as part of their clinical care (2 females and 4 males, average age of 30, ranging between 22 to 41, **Table S1**). Together, we recorded from 855 bipolar-referenced channels ('electrodes') (**Fig. 1A**). These channels spanned a total of 33 brain areas across both hemispheres (**Table S2**). Channels that displayed low signal-to-noise or frequent epileptiform discharges were excluded (Methods). For all recordings, LFPs were filtered and transformed to envelopes at alpha (8-13 Hz), beta (13-30), low gamma (30-55), mid-gamma (70-110), and high gamma (130-170) frequency bands.

During recordings, the participants engaged in free-flowing conversation with an experimenter for approximately an hour (range 20 to 92 minutes; Methods). These conversations ranged broadly in topic and theme, allowing individuals to both listen and speak (**Table S3**). All transcribed words were synchronized to neural activities at millisecond resolution. These included 3690 ± 1800 (mean ± s.t.d) words from 463 ± 230 sentences during production and 5333 ± 3038 words from 612 ± 282 sentences during comprehension (**Fig. 1b**). There were an average of 390 transitions between listening and speaking; together reflecting the dynamic interchange between the individuals involved. First order autocorrelation in both neuronal activities and the NLP model embeddings were removed (**Fig. 2a**, Method).

NLP model-based activity patterns within the cerebral network

To quantify the degree to which neural activity reflected the information being conveyed during dialogue (e.g., rather than simply to any changes in activity when speaking or listening), we employed a pre-trained GPT-2 (small) model capable of capturing the variance of brain activity in relation to linguistic information being conveyed (*35-37*). Here, the model embeddings (a set of hierarchically organized vectors serving as the artificial 'neural' activities) were trained to represent linguistic features extracted from vast language corpora and therefore allowed the degree to which neural responses reflected linguistic information conveyed during dialogue to be formally quantified (*32, 38*) (**Fig. 2a** *bottom*). Thus, for example, if the pattern of neural activity in a particular brain area consistently matched the activity derived from an artificial neural activity in these language models, this would mean that their patterns carry linguistic information about the conversations being relayed (see Methods and further below). Specifically, we used the same words that participants spoke or listened to as input to the artificial model and examined the brain-artificial neural correlations across words to manifest mesoscale neural activities that participated in language processing.

By tracking LFP signals across broadly varying conversations, we found that changes in activity in widespread parts of the brain that were consistently aligned with those of the NLP model (Methods). When accounting for all recording contacts and frequency bands, there was an overall wide distribution of correlated electrodes across most brain areas and frequency bands with the proportion of these electrodes being significantly higher than expected from chance ($p = 1.4 \times 10^{-185}$, Chi-square proportion test). The mean and standard deviation of R-value of all electrodes that displayed significant correlation was 0.09 ± 0.03 (**Fig. 2b,c**). Overall, there were more electrodes with correlated activity from areas in the left hemisphere when compared to that in the right (Chi-square test, statistic = 59, $p = 1.8 \times 10^{-14}$, **Fig. 2b, S2**). Specifically, 11% of contacts in the left hemisphere had significantly correlated activity with the NLP embeddings at the mid-gamma frequency band before word perceived, and a similar ratio of contacts had correlated activity to NLP embeddings during word production planning (Prior to word articulated: 10% in temporal and 11% in frontal areas at mid-gamma. A ratio of NLP-correlated electrode on detailed brain areas is plotted in **Fig. S2**). The highest ratio of electrodes with correlated activity was found in the left superior frontal cortex (24%; **Fig. 2c, S3**). In addition, all frontal, temporal and mesial brain areas show similar frequency-dependency (**Fig. 2d**, one-way ANOVA, p = 0.98 and p = 0.67 for production and comprehension, **Fig. S1**). These results and the general distribution of these correlated electrodes was largely consistent across participants (**Fig. S5**). Together, these findings suggest that neural activities within parts of the

frontotemporal network paralleled that of the NLP models, with similar neural representations of linguistic information captured by the models during speech production and comprehension.

To further ensure that the relationship between neural activity and those of the NLP model was indeed language-related, we randomly permutated the neural activities over words to eliminate any linguistic information obtained by chance (Methods). Through this, we found a significantly lower proportion of contacts that displayed selectivity (Chi-square test, statistic = 998, $p = 10^{-219}$, **Fig. 3a**) and a significant drop in the degree of correlation from R = 0.09 to R = 0.0. We also correlated the model embeddings to neuronal activities obtained from periods of silence of the same duration in which no conversation was held. Therefore, if correlations were simply explained by any neural activity pattern independently of language related responses, then we would expect to observe similar results. Here, we again found a markedly lower proportion of areas that displayed selectivity (Chi-square test, statistic = 399, $p = 6.9 \times 10^{-89}$), arguing that these response patterns reflected the linguistic information being conveyed.

Finally, to confirm that the above findings were generalizable, we compared the neural activity patterns to those of another NLP model trained on a different dataset. Here, we used a BERT (base) model with a bidirectional network architecture and trained on different linguistic materials (*39*) (Methods). Again, we found that the proportion of electrodes with correlated activity was significantly higher than chance (Chi-square test, statistic = 3028, $p < 10^{-100}$) and was similar or higher than that observed with the GPT-2 model (**Fig. S4**). Together, these findings suggest that the relationships between neural activity patterns in the brain and those of the language models are generalizable properties of neuronal responses.

Neural activity patterns during speech production versus comprehension

Following the above findings, we next examined whether the correlation between neural activity and the NLP models were specific to the two main components of conversation – production and comprehension. Overall, we found a similar number of electrodes whose neural activities were correlated with that of the NLP models during speech planning compared to comprehension (65 and 77 contacts, respectively; **Fig. S6**) as well as a similar strength of correlation with the models (R = 0.096 vs. R = 0.082, **Fig. 3a**). There were 10 electrodes that showed shared responses for both speech planning and comprehension, displaying certain overlapping processing between the two activities. However, most electrodes that displayed significant correlation during production did not during comprehension (Chi-square proportion test, statistics = 1.1, p = 0.30). By tagging electrodes based on the frequency band with the highest absolute correlation coefficient (Methods), we found that the best-matched frequency bands shifted towards higher frequencies for comprehension as opposed to production (**Fig. 3b**). Correlations to neural activities were reduced in alpha and beta bands after the onset of word utterance (Chi-square proportion test, statistic = 3.4, p = 0.06), while the best correlation stayed in the mid-gamma band for comprehension (Chi-square proportion test, statistic = 1.3, p = 0.24). This decrease of correlated low frequency bands was not localized in any specific areas, suggesting that it was not simply explained by outliers (**Fig. 3b**). Together, these findings reveal a general presence of neural-to-model correlations at a similar degree in both language production and comprehension. However, at the meso-scale where recordings were obtained, there were largely distinct locations and oscillatory bands that displayed diverse properties during word articulation and perception.

Neural-to-model correlations during conversation were also largely confined to distinct layers within the language models (**Fig. 3c**). Whereas lower (input) network layers within the NLP model preferentially reflect information about individual words independent of their context, higher (output) layers reflect integrated compositional sentence-level information (*40-44*). Therefore, to further examine the relationship between neural activity patterns and that of the NLP model, we 'mapped' the activities of each electrode contact onto an individual node from a network layer based on the maximum absolute correlation coefficient (Methods). Here, we found that neural activities preferentially aligned onto the high network layers for both speech planning and comprehension (21 out of 65 electrodes (32%) for production planning; 16 out of 77 (21%) for comprehension, **Fig. 3c**). For production, however, immediately following the onset of word utterance, there was a significant decrease of the number of electrodes mapping to high layers (Chi-square proportion test considering layers 10 to 12 (high), statistic = 4.9, $p = 0.027$ prior to vs. post word onset during production). These changes were broadly observed throughout the brain (**Fig. 3c**). Together, these findings therefore suggested that these activity patterns likely reflected higher order contextual information but were subserved by distinct neural mechanisms for production and comprehension.

Neural activity during turns correlated with that in response to the NLP model

Given the distinct neural activities underlying speech production and comprehension, we examined the neural processes during transitions between speaker and listener to ask whether these neural activity patterns aligned with specific transitions rather than simply any established production or comprehension. To start, we firstly tracked specific transition points at which time the participants changed from listener-to-speaker or from speaker-to-listener during dialogue. These were the times when they specifically shifted between assimilating and communicating information (**Fig. 4a**) (*20*). Across all brain areas, we found that 13% of electrodes displayed a significant change in activity during transitions from comprehension to articulation, whereas 9% displayed significant changes at the transitions to comprehension (**Fig. 4b**). Both proportions were significantly higher than random chance (Chi-square proportion test, $p = 2.0 \times 10^{-38}$ and $p = 3.0 \times 10^{-13}$). In addition, the number of contacts at the transitions from production to comprehension was significantly smaller than the transition from comprehension to production at low frequency bands (alpha, beta, low gamma bands, permutation test, $p = 7 \times 10^{-4}$, 7×10^{-4}, 0.001 respectively), whereas there were little differences at the mid- and high gamma bands ($p = 0.05$ and $p = 0.23$) between the two directions. These patterns were consistently found throughout the brain (**Fig. S7**). However, the patterns also differed in polarity in terms of increases to decreases in activity. For example, there was a decrease in gamma frequency activity in the hippocampus before transition to production, whereas electrodes from middle temporal cortices demonstrated an increase (**Fig. 4c**). Together, neural activities reflected transitions during turn-taking, showing diverse changes among brain areas and frequency bands.

Finally, given the above findings, we asked whether neural responses to these speaker-listener transitions overlapped with those reflecting correlations with the NLP models. It is possible for instance that neural responses to these transitions simply reflected differences in saliency or attention rather than the specific information being conveyed during conversation. We found, however, that 79% (61 out of 77) of the contacts whose activities correlated with that of the NLP

models overlapped with those that displayed speaker-to-listener transitions (**Fig. 4e**), while 62% (40 out of 65) of the production-planning contacts displayed changes during listener-to-speaker transitions. This degree of overlap was markedly higher than that expected by chance (Chi-square contingency; $p = 5.8 \times 10^{-18}$ and 5.3×10^{-52} for each transition direction respectively). These neural activity patterns therefore appeared to reflect not only information of the conversations being held but also tracked the specific transitions between speakers at which times the 'directionality' of communication changed – elements that together would be necessary for our ability to converse.

Discussion

Major challenges in understanding how our brain supports natural conversation include limitations in precisely recording the relevant neural activity, the difficulty of richly and accurately representing linguistic information and the ability to place both of these problems in the context of natural human interactions (*45, 46*). By comparing intracranially recorded neural patterns during natural dialogue between individuals to those of NLP models, we have tried to overcome these significant hurdles. We found that linguistic information as captured by the models could similarly explain neural activity patterns found in the brain when the participants were not only listening, but also during language production. These findings extended the previous demonstration of neural-artificial convergence focusing only on comprehension (*32, 47*) and revealed a shared neural language representation with speech production. In addition, the correlation between neural activity patterns and that of the NLP model's, by contrast, were largely absent during periods of silence or when randomizing the neural data. They were also significantly reduced after the onset of word utterance during production, together suggesting that these neural patterns reflected the specific linguistic information being conveyed.

The information being conveyed in speech production, though, did not simply mirror that of comprehension. Here, we found that neural activity patterns that reflected linguistic information tended to involve the alpha and beta frequency bands during production whereas the mid-gamma band was most salient during comprehension, suggesting the likely existence of different processing mechanisms. In addition, most of neural activities throughout frontal and temporal lobes reflected either production or comprehension; rarely were single electrodes involved in both. Together, these findings suggest a distributed computation in language processing with some mechanistic overlap between comprehension and production but substantial spatio-temporal differences as well. As a result, when tracking neural responses during comprehension-production transitions, we observed that many brain areas significantly changed their activity during turn-taking; a process surprisingly similar to that observed in prior animal and human models of communication (*20, 48*). More notably, these neural changes closely overlapped with brain areas that displayed linguistic information as identified by neural-to-model correlations, together suggesting that these response patterns reflected the process of communication rather than simply the act of listening or speaking.

The final striking finding was the relationship between neural activities and the activities of specific nodes in the NLP models (*32, 47, 49*). Overall, we found that the neural dynamics of most areas preferably mapped onto the middle and high layers of the NLP model, suggesting that these activity patterns reflected high-level information integration over the course of

conversation rather than simply related to individual words which would be represented by lower layers. Consistent with this interpretation, we found that neural-to-model correlations decreased at high layers immediately after the onset of word production, suggesting that neural responses transiently reduced in higher order contextual processing during articulation. These findings were also consistent with the proposed roles of high- and low-frequency oscillations in cognitive processes whereby lower frequency bands preferentially carried top-down propagation (*50-52*). Indeed, before word utterance, we found that neural activities reflected high-layer contextual information via lower frequencies, and we proposed that this implied a top-down planning process prior to utterance. Neural activities during speech perception also closely correlated with high NLP layers at high frequencies, likely showing a bottom-up integration of language comprehension.

Taken together, using a combination of intracranial recordings, NLP models and naturalistic conversation, our findings reveal a set of collective neural processes that could support conversation in humans and a detailed organization of neural patterns that could allow linguistic information to be shared across speakers. These observations provide a prospective neural framework by which to understand the detailed neuronal computation underlying verbal communication in humans.

Data availability

Data will be made publicly available on public repositories when the manuscript is accepted for publication.

Code availability

Codes will be made publicly available on GitHub when the manuscript is accepted for publication.

Acknowledgements

J.C is supported by the American Association of University Women, Z.M.W. is supported by NIH R01DC019653 and NIH U01NS123130 and S.S.C., A.E.H., A.C.P. are supported by NIH U01NS098968.

Author contributions

S.S.C and A.E.H. designed the study, J.C. and Z.M.W. conceived and designed the language processing analysis platform and its implementation, A.E.H and A.C.P were instrumental in data collection and pre-processing and J.C., Z.M.W. and S.S.C. performed analysis and modeling and prepared the manuscript. All authors edited the manuscript.

Competing interests

The authors declare no competing interests.

Figures

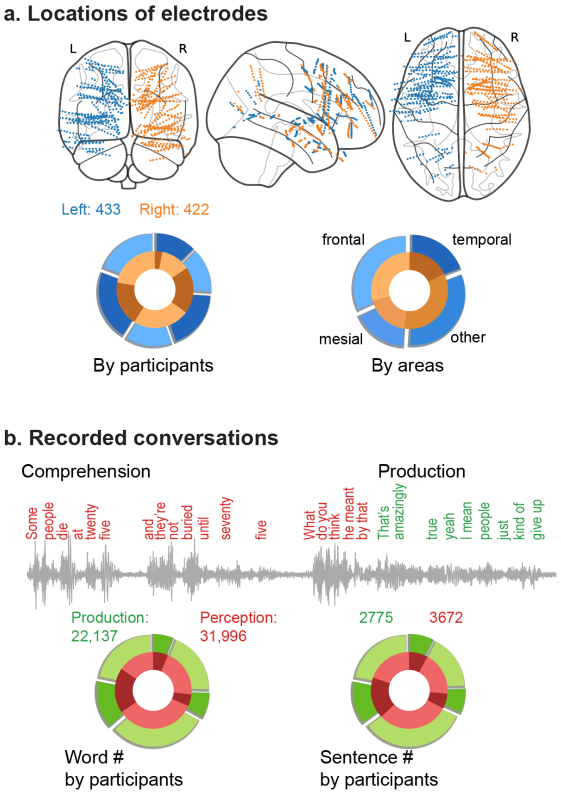

Fig. 1. Electrode distributions and speech transcription. a. Depth electrodes were implanted to six participants to record local field potentials (LFP) at 33 brain areas. A total of 855 bipolar-referenced electrode channels that demonstrated low rate of epileptiform discharges were majorly distributed at the frontal, temporal and mesial areas in both hemispheres. **b.** During voltage recordings, participants were engaged in one-to-one natural conversations with an experimenter. These conversation materials were transcribed to words, and whether a word was perceived or produced by the participant was labelled. This resulted in a total of 54,133 words together among all participants.

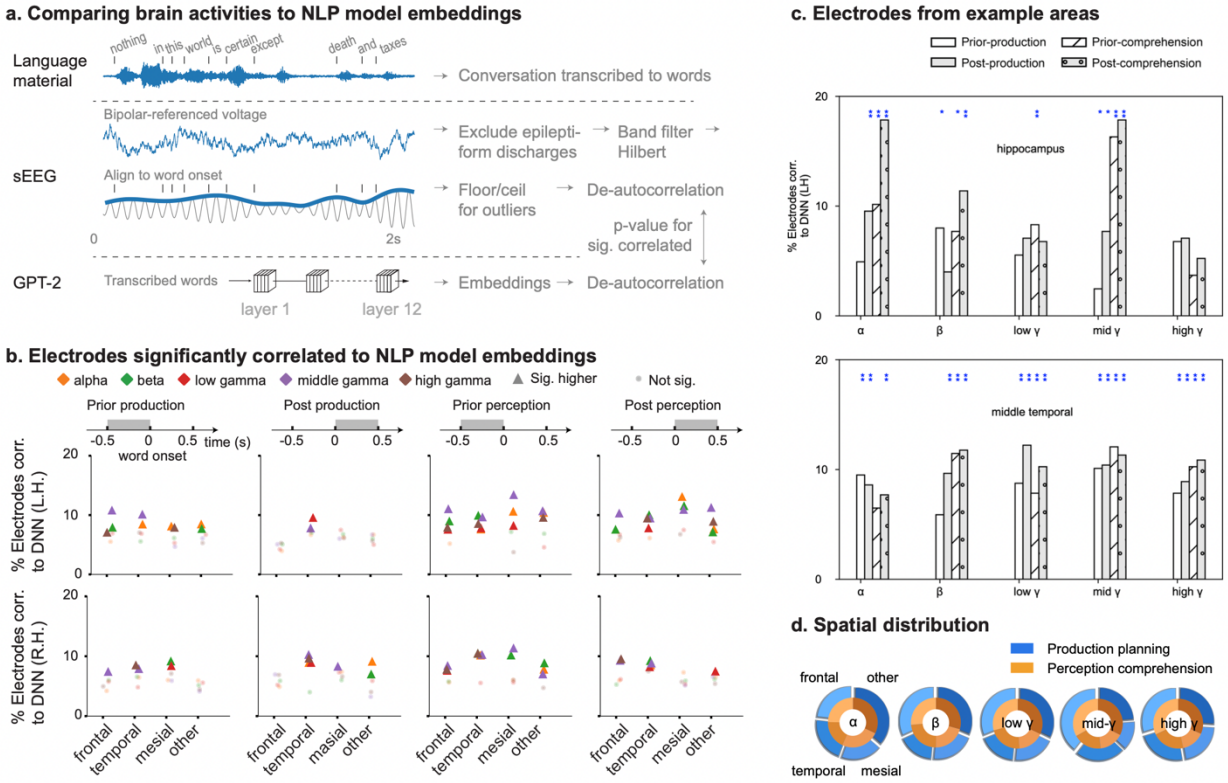

Fig. 2. Correlating brain activities to a pre-trained natural language processing model. a. Transcribed words were input to a pre-trained GPT-2 model, and the de-autocorrelated embeddings were correlated to the de-autocorrelated LFPs at each frequency band. Here, we quantified the degree to which the artificial activities of specific nodes in the language model were predictive of neural activities obtained through depth electrode recordings. Bipolar-referenced LFPs were synchronized with the auditory language materials and processed for amplitudes at the frequency bands of alpha (8-13 Hz), beta (13-30), low gamma (30-55), mid-gamma (70-110), high gamma (130-170). **b.** Across all areas, the percentage of electrodes that were significantly correlated to the artificial network embeddings is plotted. The emphasized triangles show the frequency bands and brain regions that contained a higher-than-chance number of correlated electrodes, and the rest are plotted with transparent dots. The percentage of significant channels were spreading among most of the areas in the left-hemisphere at various frequency bands for both production and perception, whereas their right-hemisphere counterparts showed a significantly lower number of correlated channels. **c.** The percentage of correlated electrodes from left hippocampus and middle temporal cortex is plotted respectively. Blue stars label the significance compared to chance level: one star: $p < 0.05$, two stars: $p < 0.01$. **d.** A similar ratio of correlated channels distributes across brain areas regardless the frequency bands.

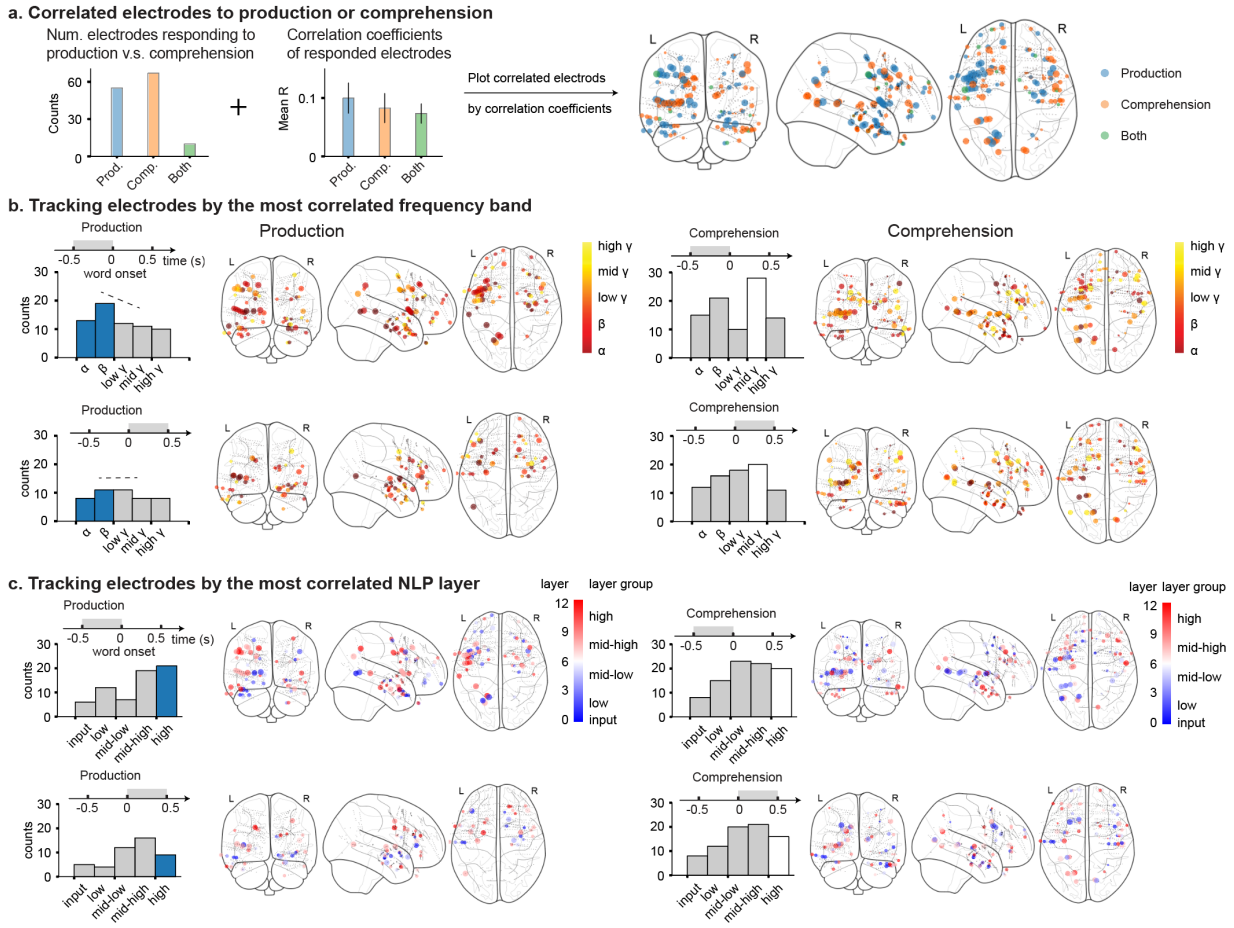

Fig. 3. Comparing neural activities during language production and comprehension. a. Demonstration of electrodes selective to artificial embeddings during one or both production and perception (*left*). Correlation coefficients are plotted by electrodes significantly correlated to artificial embeddings during perception or production (*middle*) and these electrodes by their locations (*right*). **b.** Among each electrode that significantly correlated to the artificial model, we attributed the frequency band that had the highest absolute R-value to the electrode. This 'mapping' reveals a decrease of low frequency counts after the onset of word articulation, whereas no significant changes observed at mid-gamma frequency during comprehension. **c.** Mapping the most correlated NLP layer (where the most-correlated node located) to electrodes reveals a reduce of high layer counts after the onset of word articulation, whereas the layers remain largely the same after a word is perceived. For all subplots, the size of each contact illustrates the absolute R-value of the best-matched feature.

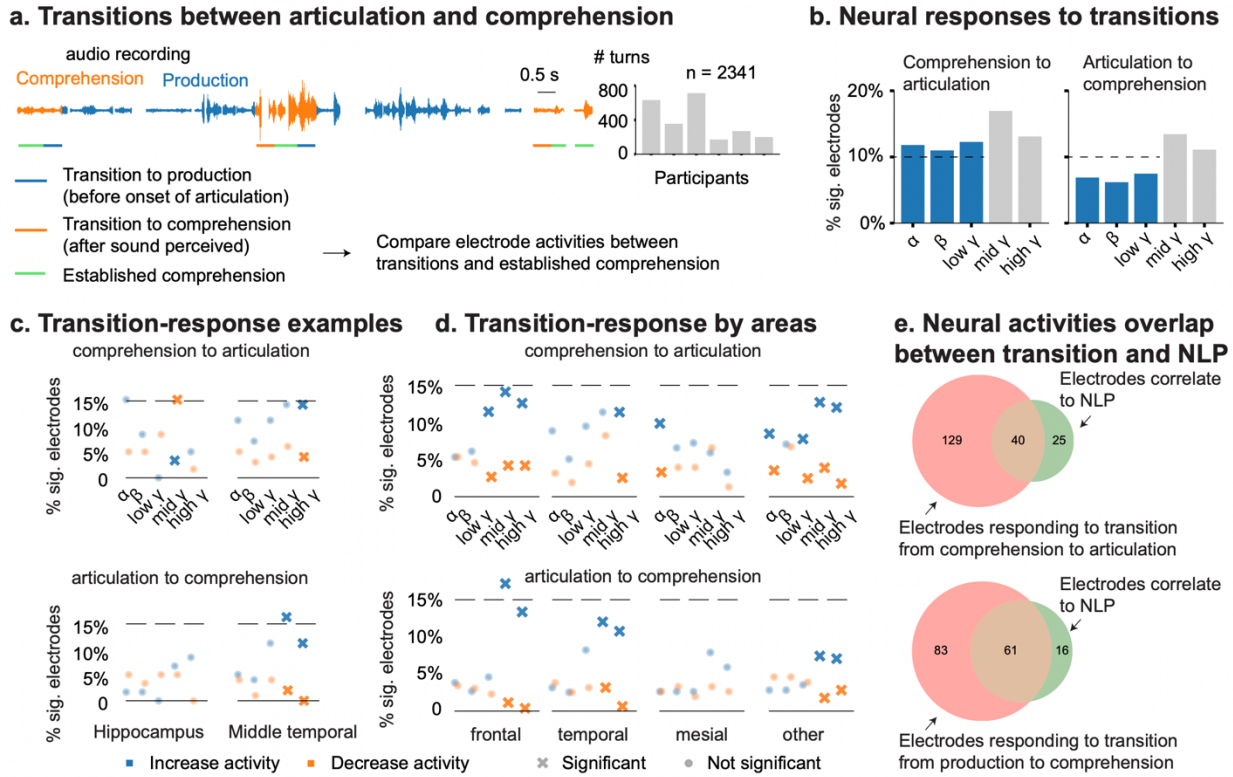

Fig. 4. Neuronal activity changes with turn-takings during conversation. a. Audio waveform illustrates the transitions between articulation and comprehension (*left*) and the counts for turn number for each participant (*right*). **b**. There are an average of 13% and 9% electrodes that showed significant changes of activities across all five bands during transitions, but transition to comprehension adopts a lower number of electrodes at low frequency bands (alpha, beta, low gamma bands) compared to transition to production, whereas the mid- and high gamma bands are similar. **c.** Percentage of electrodes that show increased activity (blue) and decreased activity (orange) are dependent on the brain area and the frequency band. The significance for each area is calculated by comparing the percentage of electrodes that show increased activities compared to that show decreased activities. **d.** Same as **c** but plotting by area groups. **e.** Comparing these electrodes to the ones that are significantly correlated to NLP during production, there is significant overlapping between NLP-response electrodes and the transition-response electrodes.

References

1. C. Brodbeck, L. Pylkkanen, Language in context: Characterizing the comprehension of referential expressions with MEG. *Neuroimage* **147**, 447-460 (2017).
2. J. Xu, S. Kemeny, G. Park, C. Frattali, A. Braun, Language in context: emergent features of word, sentence, and narrative comprehension. *Neuroimage* **25**, 1002-1015 (2005).
3. P. Hagoort, L. Hald, M. Bastiaansen, K. M. Petersson, Integration of word meaning and world knowledge in language comprehension. *Science* **304**, 438-441 (2004).
4. S. Heim, K. Alter, Prosodic pitch accents in language comprehension and production: ERP data and acoustic analyses. *Acta Neurobiol Exp (Wars)* **66**, 55-68 (2006).
5. A. Droge, J. Fleischer, M. Schlesewsky, I. Bornkessel-Schlesewsky, Neural mechanisms of sentence comprehension based on predictive processes and decision certainty: Electrophysiological evidence from non-canonical linearizations in a flexible word order language. *Brain Res* **1633**, 149-166 (2016).
6. V. Boulenger, M. Hoen, C. Jacquier, F. Meunier, Interplay between acoustic/phonetic and semantic processes during spoken sentence comprehension: An ERP study. *Brain and language* **116**, 51-63 (2011).
7. F. Bai, A. S. Meyer, A. E. Martin, Neural dynamics differentially encode phrases and sentences during spoken language comprehension. *PLoS Biol* **20**, e3001713 (2022).
8. J. W. Bohland, F. H. Guenther, An fMRI investigation of syllable sequence production. *Neuroimage* **32**, 821-841 (2006).
9. D. K. Lee *et al.*, Neural encoding and production of functional morphemes in the posterior temporal lobe. *Nat Commun* **9**, 1877 (2018).
10. S. C. Levinson, F. Torreira, Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology* **6**, 731 (2015).
11. T. Stivers *et al.*, Universals and cultural variation in turn-taking in conversation. *Proc Natl Acad Sci U S A* **106**, 10587-10592 (2009).
12. E. Fedorenko *et al.*, Neural correlate of the construction of sentence meaning. *Proc Natl Acad Sci U S A* **113**, E6256-E6262 (2016).
13. M. J. Nelson *et al.*, Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc Natl Acad Sci U S A* **114**, E3669-E3678 (2017).
14. E. Fedorenko, A. Nieto-Castanon, N. Kanwisher, Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* **50**, 499-513 (2012).
15. C. Humphries, J. R. Binder, D. A. Medler, E. Liebenthal, Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J Cogn Neurosci* **18**, 665-679 (2006).
16. L. J. Silbert, C. J. Honey, E. Simony, D. Poeppel, U. Hasson, Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proc Natl Acad Sci U S A* **111**, E4687-4696 (2014).
17. S. Heim, A. D. Friederici, Phonological processing in language production: time course of brain activity. *Neuroreport* **14**, 2031-2033 (2003).
18. G. Jang *et al.*, Everyday conversation requires cognitive inference: neural bases of comprehending implicated meanings in conversations. *Neuroimage* **81**, 61-72 (2013).
19. W. Feng *et al.*, Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain Lang* **171**, 52-61 (2017).
20. G. A. Castellucci, C. K. Kovach, M. A. Howard, 3rd, J. D. W. Greenlee, M. A. Long, A speech planning network for interactive language use. *Nature* **602**, 117-122 (2022).
21. A. K. Kuhlen, C. Allefeld, S. Anders, J.-D. Haynes, Towards a multi-brain perspective on communication in dialogue. *Cognitive neuroscience of natural language use*, 182-200 (2015).
22. S. Bögels, S. C. Levinson, The brain behind the response: Insights into turn-taking in conversation from neuroimaging. *Research on Language and Social Interaction* **50**, 71-89 (2017).

23. S. A. Ghilzai, M. Baloch, Conversational analysis of turn taking behavior and gender differences in multimodal conversation. *European Academic Research* **3**, 10100-10116 (2015).
24. S. Bögels, L. Magyari, S. C. Levinson, Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific reports* **5**, 1-11 (2015).
25. F. Ferreira, B. Swets, How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language* **46**, 57-84 (2002).
26. openai. (openai.com, 2022).
27. Y. Zhang *et al.*, Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, (2019).
28. T. Wolf, V. Sanh, J. Chaumond, C. Delangue, Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, (2019).
29. P. Budzianowski, I. Vulić, Hello, it's GPT-2--how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*, (2019).
30. J. Vig, Y. Belinkov, Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, (2019).
31. Y. Belinkov, J. Glass, Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics* **7**, 49-72 (2019).
32. A. Goldstein *et al.*, Shared computational principles for language processing in humans and deep language models. *Nat Neurosci* **25**, 369-380 (2022).
33. C. Caucheteux, J.-R. King, Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.2007. 2003.186288 (2021).
34. A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98**, 630-644 e616 (2018).
35. A. Radford *et al.*, Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
36. A. Vaswani *et al.*, Attention is all you need. *Advances in neural information processing systems* **30**, (2017).
37. T. Wolf *et al.*, Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, (2019).
38. M. Schrimpf *et al.*, The neural architecture of language: Integrative modeling converges on predictive processing. *Proc Natl Acad Sci U S A* **118**, (2021).
39. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, (2018).
40. R. Futrell, E. Gibson, R. P. Levy, Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cogn Sci* **44**, e12814 (2020).
41. Q. T. Ho, T. T. Nguyen, N. Q. Khanh Le, Y. Y. Ou, FAD-BERT: Improved prediction of FAD binding sites using pre-training of deep bidirectional transformers. *Comput Biol Med* **131**, 104258 (2021).
42. K. Huang, A. Hussain, Q.-F. Wang, R. Zhang, in *Cognitive Computation Trends,*. (Springer International Publishing : Imprint: Springer,, Cham, 2019), pp. 1 online resource (VII, 163 pages 166 illustrations, 146 illustrations in color.
43. K. R. Chowdhary. (Springer (India) Private Limited,, New Delhi, 2020), pp. 1 online resource (730 p.
44. P. Koehn. (Cambridge University Press,, New York, 2020), pp. 1 online resource.
45. Y. Hmamouche, L. Prevot, O. Magalie, C. Thierry, in *Interspeech 2020*. (ISCA, 2020), pp. 101-105.
46. J. J. Van Berkum, M. Spivey, K. McRae, M. Joanisse, The electrophysiology of discourse and conversation. (2012).

47. M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, F. P. De Lange, A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* **119**, e2201968119 (2022).
48. D. E. Okobi Jr, A. Banerjee, A. M. Matheson, S. M. Phelps, M. A. Long, Motor cortical control of vocal interaction in neotropical singing mice. *Science* **363**, 983-988 (2019).
49. M. Schrimpf *et al.*, Artificial neural networks accurately predict language processing in the brain. *BioRxiv* **10**, 26.174482 (2020).
50. T. J. Buschman, E. K. Miller, Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *science* **315**, 1860-1862 (2007).
51. C. G. Richter, W. H. Thompson, C. A. Bosman, P. Fries, Top-down beta enhances bottom-up gamma. *Journal of Neuroscience* **37**, 6698-6711 (2017).
52. J. R. Iversen, B. H. Repp, A. D. Patel, Top-down control of rhythm perception modulates early auditory responses. *Annals of the New York Academy of Sciences* **1169**, 58-73 (2009).

MATERIALS AND METHODS

In this study, all data acquisition and analyses were approved by the Partners Human Research Committee Institutional Review Board (currently Massachusetts General Brigham Institutional Review Board). Before their enrolment in the study, participants were informed that their participation would not impact their clinical treatment, and that they could withdraw their participation at any time without impact of their clinical care. Consent was obtained from all study participants.

Stereo EEG recording

We recorded neurophysiological data from six intractable epilepsy participants when they went through invasive monitoring for localization of the brain areas that onset their seizures (**Table 1**). A team of clinicians who were independent of this study determined the decision to implant electrodes, as well as the types of electrodes and the areas for implantation solely based on clinical grounds. The reference of all the voltage recordings was to one EEG contact.

We recorded neuronal activities using a neural signal processor system (128 to 256 channels, NSP, Cerebus, Blackrock Microsystems). These raw voltages were digitized at a 2 kHz resolution and then filtered online to capture local field potentials (low-pass filter, 1 kHz cutoff). In total, we obtained recordings from 1154 electrodes from six participants.

Speech recordings

We recorded the speech perception and production of the participants from conversations between the participant and the experimenter (> 20 minutes). These conversations varied broadly in topic and theme to allow the participants to actively engage in dialogue whereby they both listened and spoke (**Table S3**). For example, conversations would involve real-world topics such as a movie that was recently watched or interpersonal thoughts and events shared between the experimenter and the subject.

We used an audio recorder (DR-40X by TASCAM) to record the conversation. The audios were synchronized with the intracranial activity at a millisecond resolution by adding an analog input channel to the NSP (Blackrock Microsystems). Audio recordings for each participant were manually transcribed to yield a time stamp for each word as well as the speaker identity (participant or experimenter) using SpeechScribe, a Python application developed in house for manual speech annotation. Any words or sentences containing participants' personal information were removed and replaced by a set of other names and digits. Recordings in which the participant did not interact with another person were used for control comparison.

Electrode localization

We adopted a combined surface registration and volumetric process to identify each electrode's anatomical location within the brain (*1, 2*). We first used FreeSurfer (*3*) to align preoperative MRI data with a postoperative CT scan, then we manually transformed electrode locations identified from the CT scan into the MRI space (*4*). We mapped each electrode to a set of brain regions defined in the DKT atlas using an electrode labeling algorithm (ELA) (*5-7*). The electrodes were assigned with an additionally probability of being situated in white matter. Next,

the locations of electrodes were grouped into frontal, temporal, mesial and the remaining regions ('other' category) based on their gray matter regions, and electrodes locating at white matter were combined with the 'other' category (**Table S2**).

Bipolar referencing

Neural data re-referenced into a 'bipolar' configuration by taking the voltage differences between adjacent channels from the same electrode array in MATLAB (MathWorks). The importing of neural data was assisted by the NPMK toolbox (Blackrock Microsystems). Channels that showed no variance across the whole recording period were excluded from further analysis. The voltages were further decimated to 1 kHz. In total, we obtained 1065 electrode pairs from six participants.

Interictal epileptiform discharge detection

Though no seizure occurred during the language tasks, interictal discharges may be common. In order to remove these possible large amplitudes confounds from further data analysis, we detected epileptiform discharges automatically (*8*) and removed channels in which there were a high rate of epileptiform events (> 6.5 discharges/min). This resulted in a total of 855 bipolar electrodes remaining from six participants.

Frequency bands computation and alignment to words

LFPs were further processed for amplitudes at alpha (8-13 Hz), beta (13-30), low gamma (30-55), mid-gamma (70-110), high gamma (130-170) frequencies using the scipy package for signal processing in Python with the following steps: For each bipolar-referenced voltage time series, the time stamps during interictal discharges were replaced by an interpolation of voltages 2 seconds before and after the onset of the epileptiform discharges. Then to construct the envelope of each frequency band, the voltages were filtered by a Chebyshev type II filter with 4th order and 40 dB attenuation for each frequency band, and a Hilbert transformation was further applied. Next, outliers of the envelopes were capped to $q3 + c \cdot (q3 - q1)$ and floored to $q1 - c \cdot (q3 - q1)$, where $q1$ and $q3$ are the first and third quartiles, and $c = 5$ determines the threshold of outliers.

To compute the neuronal activities for each word, envelopes were initially aligned to the auditory signals. For each word, the neuronal activities at each frequency band were computed as the average of the envelope over a 0.5 second window. Words during epileptiform discharges were removed from further analysis. We further treated the outliers with $c = 1.5$ using the same method as described in the previous paragraph to ensure that our findings were not driven by just a few points. After this step, for each frequency band of an electrode, we obtained an array of neural activities at the same dimension as the word number.

De-autocorrelation

Since the envelopes at all frequency bands were auto-correlated and the first order autocorrelation (AR(1)) dominated all higher orders, we performed further procedure to remove the first-order autocorrelation. The first order autocorrelation of neural activities as the sequence of words was calculated for each frequency band:

$$\rho = \mathrm{corr}(X_i, X_{i-1}), \tag{1}$$

where X_i is the neuronal activity for the i-th word. Then we removed the AR(1) by transforming the envelopes by

$$\tilde{X}_i = X_i - \rho X_{i-1}. \tag{2}$$

Natural language processing (NLP) network embeddings

We selected a Generative Pre-trained Transformer (GPT-2, small) model to encode language information and compare to brain activities (*9, 10*): The OpenAI GPT-2 model has been applied in previous work to multiple brain imaging studies and has been shown capable of capturing variance of brain activities from multiple areas during language comprehension (*11-13*). This model was trained on a combination of text datasets that were built on a wide range of natural language demonstrations including diverse domains and contexts, hence, the GPT-2 model was able to process natural dialogue with informal colloquial terms (e.g., "gonna" rather than "going to") often encountered during real-world conversations (*14-17*). We directly loaded the tokenizer (mapping of a word to a vector) and the model from the pre-trained GPT-2 small model in PyTorch via huggingface without fine-tuning (*18*). The model was built on 12 structurally identical modules in sequence and input sentences were represented between layers by hidden embeddings, which were 768-dimension vectors for each word per layer (*i.e.* 768 nodes or artificial neurons). The model outputs 13 sets of hidden embeddings with 768 dimensions for each word of a sentence. These 13 embeddings include the embeddings 'output' of the 12th hidden layers and the 'input' embedding to the first module. To capture the contextual information of the discourse for each word, we input words with a moving window of 600 words (concatenation of current and previous words) to NLP model to get the embeddings. For words corresponding to more than one tokens (part of words), the embeddings were obtained from the last token of the word. After this step, we got the embeddings with the dimension of number of words x 768 nodes x 13 layers for each participant.

Correlating neural activities to NLP model embeddings

To examine how brain activities involve in language processing, we correlated neural activities to artificial embeddings by checking all neural-artificial pairs. Specifically, after de-auto-correlating the artificial embeddings following the same procedures described in Eq. (2), we fit the neural activities (as a function of words) per frequency band of an electrode using the artificial embeddings (on the same words) of a node to a linear regression, and obtained the Pearson correlation and the p-value of whether they were significantly corelated (H_0: slope = 0). Any neural-artificial pair with p-values lower than 0.05 (Bonferroni corrected for multiple comparison over 768 dimensions) was considered as correlated. Then for a given brain area, we examined whether there were significant number of electrodes correlated to the NLP model based on Chi-square test ($p < 2.5 \times 10^{-3}$, Bonferroni corrected for 13 layers).

Grouping brain areas based on left and right hemispheres

Given that some participants in this study were left-handed, we further examined the neural responses to language based on the hemispheres from participants of different handedness.

Interestingly, among the left-handed subjects, correlated electrodes from left-hemisphere were still significantly higher than those from the right hemisphere (**Fig. S3**, 2-side permutation test with combining frequency bands and layers, $p < 10^{-4}$). Hence, instead of grouping hemispheres based on the handedness, we combined all participants and separated electrodes by left and right hemisphere regardless of their handedness in **Fig. 2**.

Control with BERT (base) model as another NLP model

To ensure the observed neural-artificial correlation was generalizable across embeddings from different NLP models, we calculated the neural correlation to a pre-trained BERT (base) model (*19*). BERT (base) model was composed of 12 layers with similar architecture (transformer modules) to the GPT-2 model and was independently trained on different language corpus. The modules in BERT model examined pairs of words including both past and future words. Due to this bi-directional pairing design, we separated language materials by sentences and obtained the embeddings of the words by inputting one sentence at a time. All the other procedures were exactly the same as that of the GPT-2 model. Note that the percentage of BERT-correlated electrodes after words articulated was higher than using GPT-2 model (**Fig. S4**). This was possibly because the BERT model contained future word information from the same sentence, so the high percentage may come from articulation planning of future words in the sentence.

Controls with randomly shuffling and silence periods

To ensure that the percentage of correlated electrodes was not originated from randomness and was rooted in language processing, we performed following two procedures as controls to the neural-artificial correlation: First, we randomly permutated the neural activities over words before the calculation of correlations while keeping all other procedures intact and calculated the percentage of electrodes that were significantly correlated to the embeddings. Second, we managed to obtain silent periods of neuronal recordings from five participants when there were no speech produced or perceived, and examined the correlations between neural activities during these silent periods and the artificial embeddings during the actual conversations. Therefore, if the observed neural-artificial correlations had originated from randomness or other inherent neural properties unrelated to language processing, we would have observed similar correlations as those with actual neural activities.

Mapping electrodes to frequency band and artificial layer

Since an electrode could show significant correlations from multiple frequency bands to embeddings in multiple layers, in this section, we assigned the electrode to a certain frequency or layer by considering the maximum of the absolute R-value to find the one that optimally described the neural activity. Specifically, the significance of an electrode was determined by the threshold of 0.05 with the Bonferroni correction for multiple comparison over 768 embedding dimensions, 13 layers and 5 frequency bands. This resulted in 65 and 77 significant electrodes during speech planning and comprehension. We performed further analysis to show that these NLP-correlated activities cannot be explained by correlations to the low-level sound features because the percentage of the NLP-correlated electrodes that were significantly correlated to the amplitude of the voice were 9% and 7% lower than all electrodes respectively (2-side permutation test combined, $p = 0.025$, $n = 10000$).

For each of these electrodes, we picked the frequency band that showed the highest absolute R-value, and tagged this frequency to the electrode. In this way, we 'mapped' the frequency band to each electrode that was significantly correlated to the artificial embeddings. We plotted these electrodes with their mapped frequency bands based on the electrode positions in the brain (**Fig. 3b**). Similarly, since we examined neural correlations to all embeddings from different artificial layers, we repeated similar method by selecting the NLP layers that showed the highest absolute R-values for each electrode that was significantly correlated to the NLP, and tagged the layer to the electrode. **Fig. 3c** plotted the locations of these electrodes and colored by the layers.

Neural activities correlate to turn-taking

We investigated the turn-taking properties of a conversation by examining whether any electrodes show significant changes in each frequency envelope during speaker-listener transitions. For transitions from comprehension to articulation, we averaged neural activities at each frequency band using a 0.5 s time window before the onset of the first word articulation; For transitions to comprehension, we averaged neural activities from 0.5 s window after the onset of the first word perception. The average activities at each frequency band were compared to the activities during established comprehension using a T-test from a Scipy package (ttest_ind) and the threshold of p-value to determine whether the activities changed significantly was set to be 0.05. Because speaker-listener transitions had much longer intervals compared to a word being articulated or perceived, we did not perform de-autocorrelation for this analysis. Similar to the method in the previous section, we selected the band that showed the lowest p-values and attributed each electrode responding to transitions with the selected frequency band.

To examine whether the number of electrodes was significantly different from speaker-listener transition to listener-speaker transition, we used two neural populations with each including all electrodes and we labelled each electrode by whether or not it was significantly responding to one direction of the transitions or the other. Next, we used these populations to perform a 2-side permutation test: We concatenated the two populations and randomly sampled from the mixture for a given number of times, with each sampling contained the same number of observations as the original two populations. Then for each sampling, the absolute difference between the average values of the two sets of random sampling was calculated, and this value was compared to the absolute difference from the actual populations. After repeating these steps of sampling, the p-value was then defined as the number of times when the difference of random sampling from mixed population was greater than that from the actual population, divided by the number of drawings.

Supplementary Tables

Table S1. Summary of study participants, including gender and handedness.

Subject	Handed-ness	Gender
Participant 1	right	male
Participant 2	right	male
Participant 3	left	female
Participant 4	left	male
Participant 5	left	male
Participant 6	right	female

Table S2. Electrode distribution across brain areas

Area group	Area	Elec. # left	Elec. # right
frontal	Caudal middle frontal	21	17
	Lateral orbitofrontal	21	21
	Medial orbitofrontal	2	6
	Pars orbitalis	8	10
	Pars triangularis	11	16
	Precentral	11	6
	Rostral middle frontal	45	31
	Superior frontal	19	17
mesial	Amygdala	20	15
	Caudal anterior cingulate	13	9
	Caudate	13	11
	Hippocampus	25	34
	Parahippocampal	1	5
	Rostral anterior cingulate	2	4
temporal	Fusiform	0	13
	Inferior temporal	4	6
	Middle temporal	51	47
	Superior temporal	28	9
other	Accumbens area	5	1
	Cerebral-white-matter	102	100
	Entorhinal	1	3
	Inferior parietal	2	4
	Insula	3	4
	Isthmus cingulate	3	0
	Lingual	3	0
	Pallidum	3	6
	Pars opercularis	4	0
	Pericalcarine	4	0
	Postcentral	0	3
	Putamen	7	6
	Superior parietal	0	10
	Supra marginal	0	8
	Superior parietal	1	0

Table S3a. Example 1 of the natural dialogue between experimenter (E) and participant (P).

Speaker	Sentences
E	Okay Batman would make a great Dad
P	I just saw the new Batman movie the Lego Batman movie
E	Any good
P	Yeah it's pretty good
E	I think like the thing has there been one before

Table S3b. Example 2 of the natural dialogue between experimenter (E) and participant (P).

Speaker	Sentences
E	I was gonna ask when's the last time a joke made you laugh but technically
P	Oh I laugh all the time at jokes
E	Okay
P	Even the dumb ones honestly
E	Do you have a favorite comedian or
P	I think ah not really I like Seinfeld a lot
E	Okay
P	this one guy uh Louis CK
E	Louis CK is awesome
P	He's pretty funny
E	I think he actually comes from Boston
P	Yeah he he comes from around Boston yeah

Table S3c. Example 3 of the natural dialogue between experimenter (E) and participant (P).

Speaker	Sentences
P	I watch like Top Chef uh a few times and like Hell's kitchen and stuff
E	What's Top Chef
P	Those are pretty good
E	So
P	It's fun where like they get a bunch of people together and they give them a bunch of like challenges and then they vote them off one by one
E	No okay it's like survivor style
P	Yeah basically

Supplementary Figures

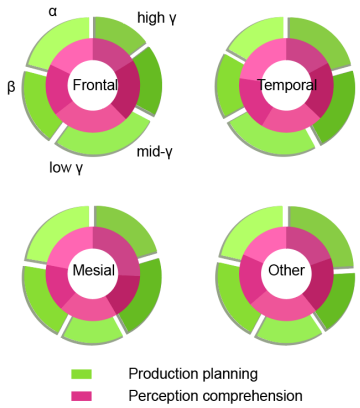

Figure S1. NLP-correlated electrodes across brain areas and frequency bands. Electrodes that were significantly correlated to NLP were grouped by frequency bands. The distribution of correlated channels across different areas remained largely the same (Repeated Measures ANOVA, $p = 0.99$ and 0.14 for production and perception).

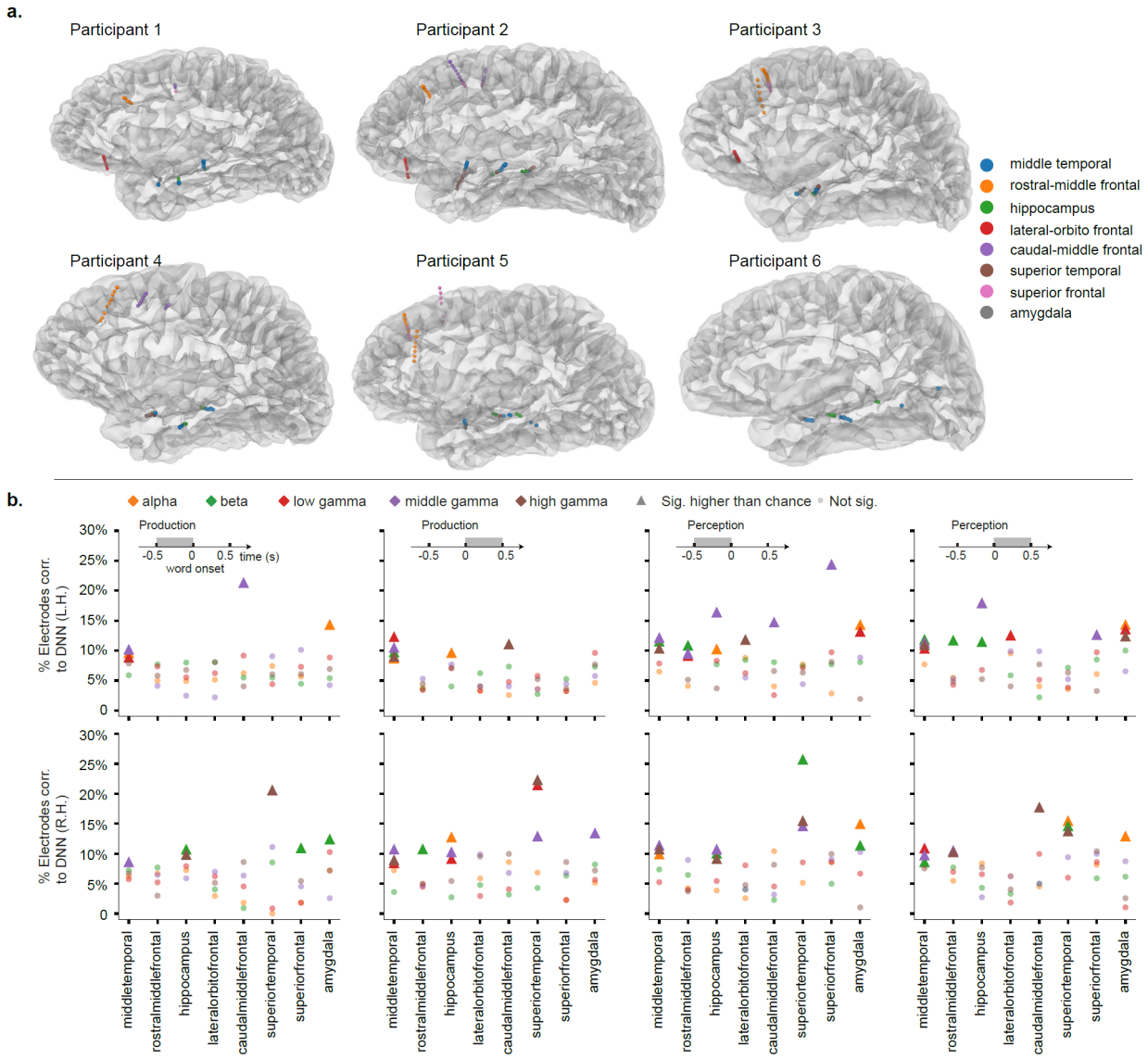

Figure S2. Artificial-neural activity correlations by granular electrode locations. a. Locations of electrodes parcellated into different granular brain areas by subjects. **b.** The percentage of electrodes that were significantly correlated to the artificial network embeddings is plotted by granular brain areas.

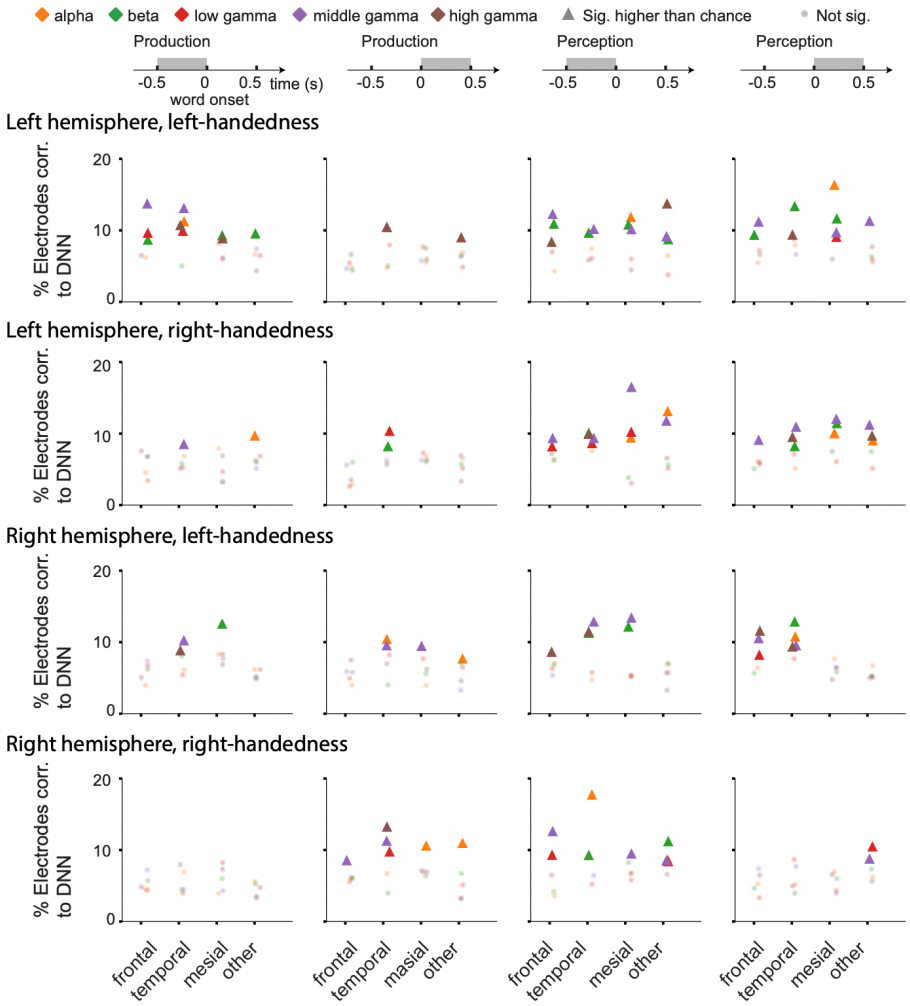

Figure S3. Artificial-neural activity correlations by brain areas grouped by hemisphere and the participants' handedness. The percentage of electrodes that were significantly correlated to the artificial network embedding is plotted. Among left-handed participants, there were higher percentage of correlated electrodes in left hemisphere compared to the right (permutation test, $p < 10^{-4}$).

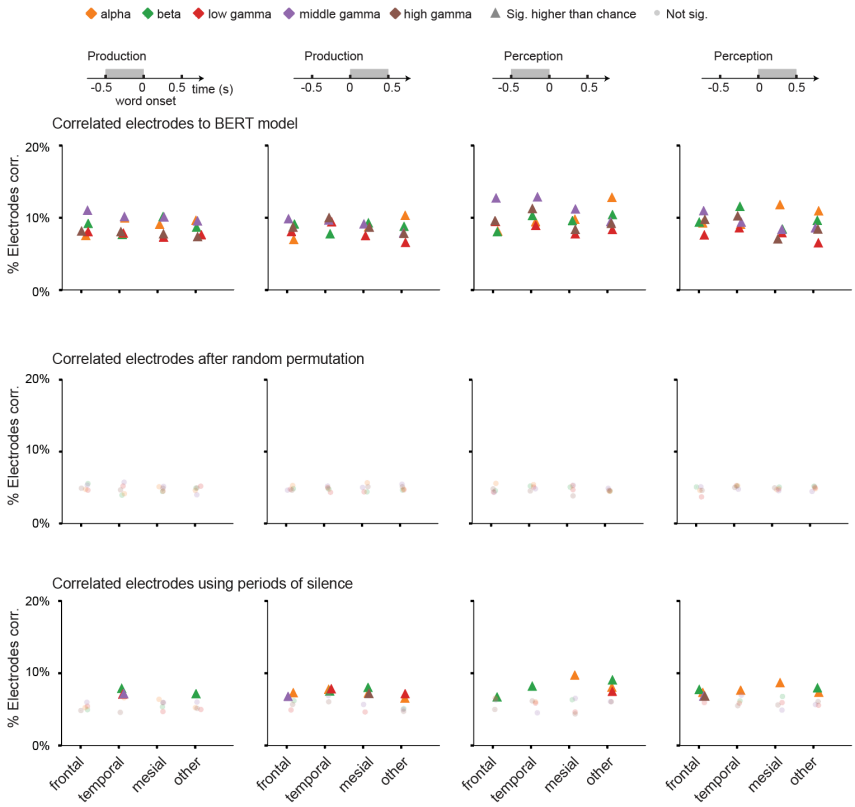

Figure S4. Controls for generalization and other factors. Percentage of correlated electrodes to another pre-trained NLP model (BERT) showed similar percentage to the pre-trained GPT-2 model, suggesting the neural-artificial convergence is generalizable (*top*). A control with the permutated NLP embeddings showed that correlated electrodes were homogeneously distributed at chance-level (5%) (*middle*), which were significantly lower than those without permutation (Chi-square test, statistic = 998, $p = 10^{-219}$). A control with LFPs obtained from a period when no speech perceived or produced showed significantly reduced the number of correlated channels (*bottom*) (Chi-square test, statistic = 399, $p = 6.9 \times 10^{-89}$).

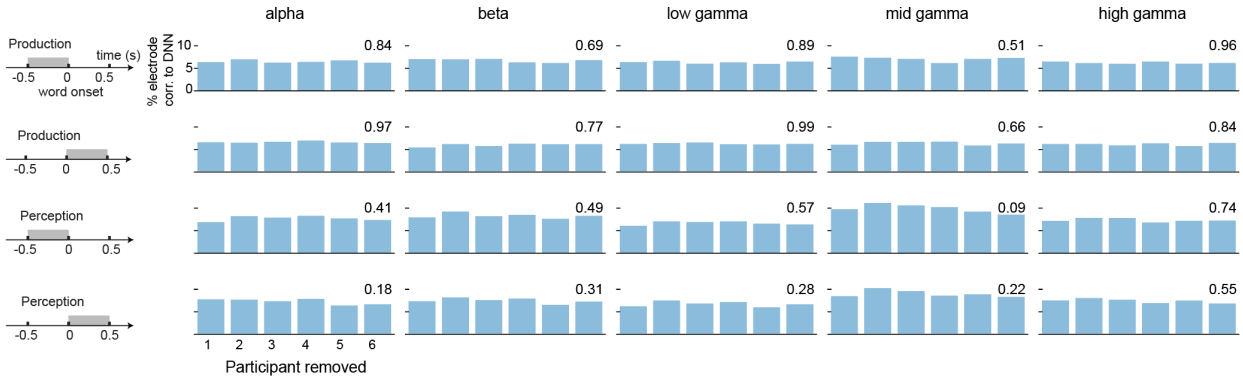

Figure S5. Percentage of electrodes correlating to NLP model embeddings by removing a participant. To ensure the ratio of electrodes significantly correlated to NLP was not caused by a single participant, we removed each participant (x-axis) and plot the percentage for both production (top 2 rows) and perception (bottom 2 rows) at different frequency bands. An one-way ANOVA was applied to examine whether removing a patient caused a significant change of the ratio, and the p-values are texted on each subplot. No removal of a participant resulted in significant change of the correlated electrode ratio.

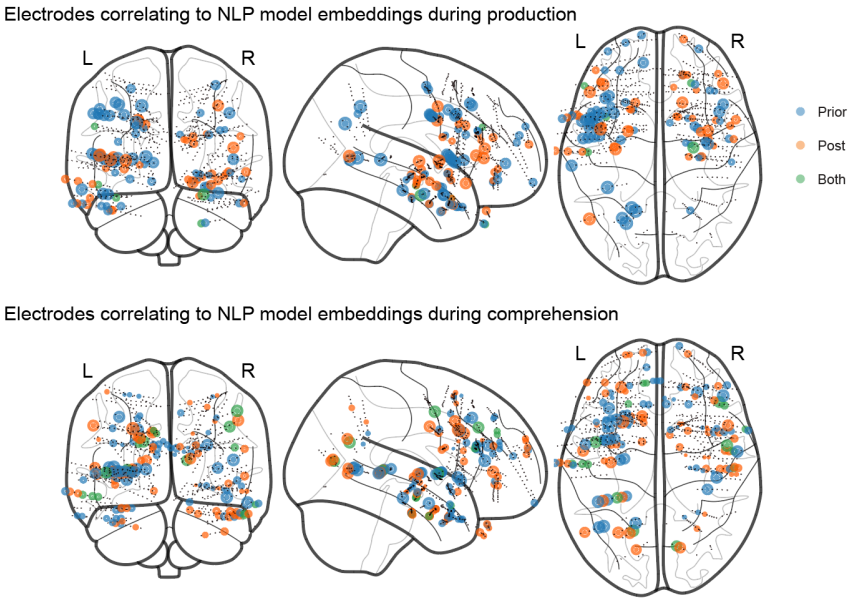

Figure S6. **Distribution of electrodes correlating to NLP model embeddings before and after word onset during speech production and comprehension.** Mapping each electrode by the time period when they were correlated to the artificial model embeddings.

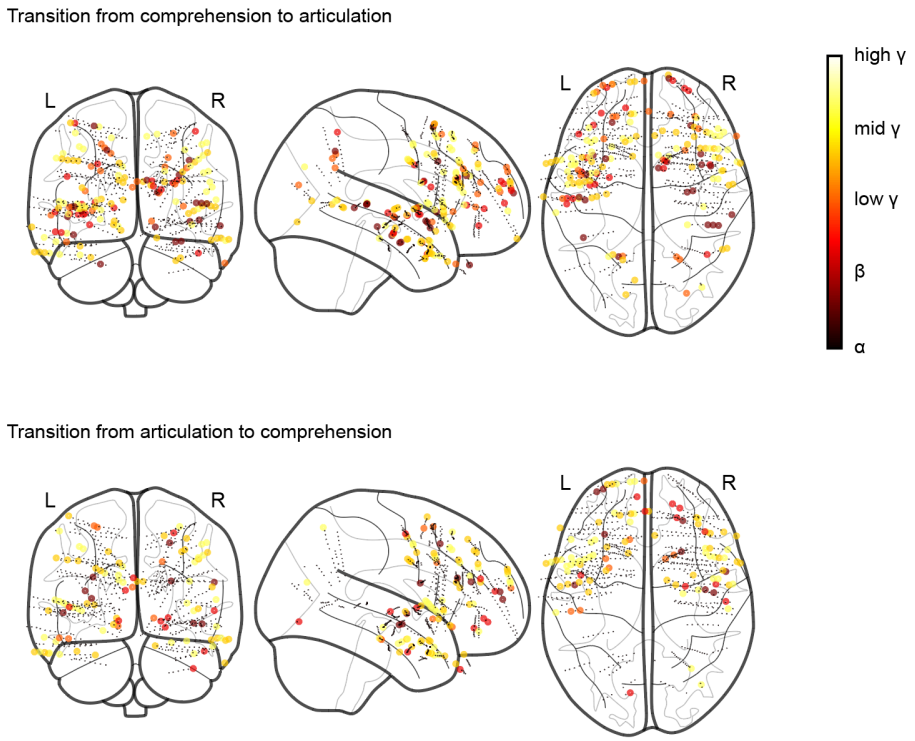

Figure S7. Electrodes with frequency bands significantly changed during speaker-listener transitions. Mapping of electrodes by their frequency bands that showed most significant changes during transitions.

References

1. G. Postelnicu, L. Zollei, B. Fischl, Combined volumetric and surface registration. *IEEE Trans Med Imaging* **28**, 508-522 (2009).
2. L. Zollei, A. Stevens, K. Huber, S. Kakunoori, B. Fischl, Improved tractography alignment using combined volumetric and surface registration. *Neuroimage* **51**, 206-213 (2010).
3. B. Fischl, FreeSurfer. *Neuroimage* **62**, 774-781 (2012).
4. A. R. Dykstra *et al.*, Individualized localization and cortical surface-based registration of intracranial electrodes. *Neuroimage* **59**, 3563-3570 (2012).
5. B. Fischl *et al.*, Automatically parcellating the human cerebral cortex. *Cerebral cortex* **14**, 11-22 (2004).
6. O. Felsenstein, N. Peled, MMVT–multi-modality visualization tool. *GitHub Repository*, (2017).
7. M. Reuter, H. D. Rosas, B. Fischl, Highly accurate inverse consistent registration: a robust approach. *Neuroimage* **53**, 1181-1196 (2010).
8. R. Janca *et al.*, Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings. *Brain Topogr* **28**, 172-183 (2015).
9. A. Radford *et al.*, Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
10. A. Vaswani *et al.*, Attention is all you need. *Advances in neural information processing systems* **30**, (2017).
11. A. Goldstein *et al.*, Shared computational principles for language processing in humans and deep language models. *Nat Neurosci* **25**, 369-380 (2022).
12. M. Schrimpf *et al.*, The neural architecture of language: Integrative modeling converges on predictive processing. *Proc Natl Acad Sci U S A* **118**, (2021).
13. C. Caucheteux, J.-R. King, Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.2007. 2003.186288 (2021).
14. P. Budzianowski, I. Vulić, Hello, it's GPT-2--how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*, (2019).
15. Z. Sun, R. Zemel, Y. Xu, A computational framework for slang generation. *Transactions of the Association for Computational Linguistics* **9**, 462-478 (2021).
16. Y. Zhang *et al.*, Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, (2019).
17. T. Wolf, V. Sanh, J. Chaumond, C. Delangue, Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, (2019).
18. T. Wolf *et al.*, Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, (2019).
19. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, (2018).