Current Biology
Report

Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech

Michael P. Broderick,[1,6,*] Andrew J. Anderson,[2] Giovanni M. Di Liberto,[1,3,4] Michael J. Crosse,[1,5] and Edmund C. Lalor[1,2,*]
[1]School of Engineering, Trinity Centre for Bioengineering, and Trinity College Institute of Neuroscience, Trinity College Dublin, College Green, Dublin 2, Ireland
[2]Department of Biomedical Engineering, Department of Neuroscience, and Del Monte Institute for Neuroscience, University of Rochester, Rochester, NY 14627, USA
[3]Laboratoire des Systèmes Perceptifs, CNRS, 29 Rue d'Ulm, Paris 75005, France
[4]Département d'Etudes Cognitives, ENS, PSL Research University, 60 Rue Mazarine, Paris 75006, France
[5]Department of Pediatrics and Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[6]Lead Contact
*Correspondence: brodermi@tcd.ie (M.P.B.), edmund_lalor@urmc.rochester.edu (E.C.L.)
https://doi.org/10.1016/j.cub.2018.01.080

SUMMARY

People routinely hear and understand speech at rates of 120–200 words per minute [1, 2]. Thus, speech comprehension must involve rapid, online neural mechanisms that process words' meanings in an approximately time-locked fashion. However, electrophysiological evidence for such time-locked processing has been lacking for continuous speech. Although valuable insights into semantic processing have been provided by the "N400 component" of the event-related potential [3–6], this literature has been dominated by paradigms using incongruous words within specially constructed sentences, with less emphasis on natural, narrative speech comprehension. Building on the discovery that cortical activity "tracks" the dynamics of running speech [7–9] and psycholinguistic work demonstrating [10–12] and modeling [13–15] how context impacts on word processing, we describe a new approach for deriving an electrophysiological correlate of natural speech comprehension. We used a computational model [16] to quantify the meaning carried by words based on how semantically dissimilar they were to their preceding context and then regressed this measure against electroencephalographic (EEG) data recorded from subjects as they listened to narrative speech. This produced a prominent negativity at a time lag of 200–600 ms on centro-parietal EEG channels, characteristics common to the N400. Applying this approach to EEG datasets involving time-reversed speech, cocktail party attention, and audiovisual speech-in-noise demonstrated that this response was very sensitive to whether or not subjects understood the speech they heard. These findings demonstrate that, when successfully comprehending natural speech, the human brain responds to the contextual semantic content of each word in a relatively time-locked fashion.

RESULTS

Electroencephalographic (EEG) data were recorded from subjects as they listened to narrative speech in the form of audiobook recordings. Importantly, and as detailed below, almost all of the data we present were acquired as part of previously published studies and not with the goal of studying semantic processing. To relate the neural data to the semantic processing of this speech, we first wished to parameterize the speech stimuli such that individual words were quantified according to their semantic context. There are many ways to do this. Inspired by the brain's sensitivity to incongruous new words (as seen in the N400 [3–6]), we chose to do it based on quantifying how "semantically dissimilar" each new word was compared to its immediately preceding context. This idea of semantic distance has previously been used in studies of reading-time effects [15], reading comprehension [17], and brain imaging of speech processing [17]. Our specific approach was based on the well-known word2vec model [16], whereby each word in a speech stimulus is converted to a high-dimensional vector (in our case 400 dimensions, which was a somewhat arbitrary choice; see STAR Methods). The idea is that words that share common contexts (i.e., co-occur) in a very large corpus of text are converted to vectors that are located in close proximity to one another in this high-dimensional space. As such, although it is more a measure of local word context and not a direct measure of semantics per se, the vector associated with each word can be used as a proxy for the word's meaning. We then defined the "semantic dissimilarity" of each specific content word by comparing (via a Pearson's correlation) its 400-dimensional vector with the average of the vectors corresponding to all the preceding words in that particular sentence—word combination by averaging has proved practically effective in previous research [18, 19]—and then subtracting that correlation from 1. Where a specific word

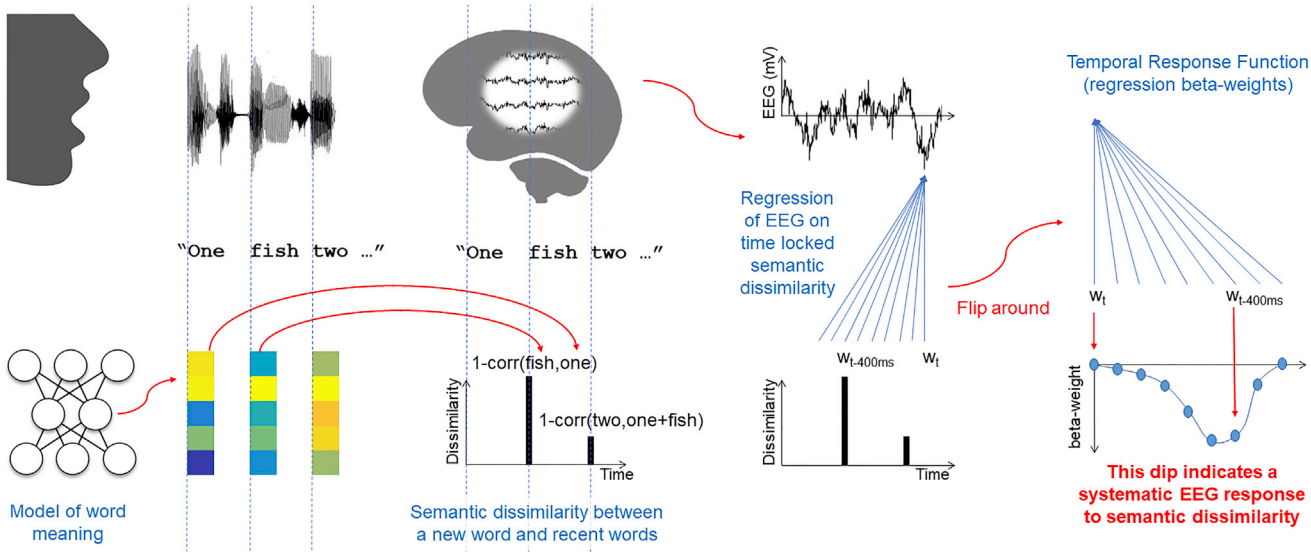

Figure 1. Regularized Regression Analysis for Estimating an Electrophysiological Correlate of Semantic Dissimilarity to Natural Speech

Content words from an audiobook are converted to 400-dimensional vectors using the word2vec algorithm [20] (bottom left). The semantic similarity of each word to its preceding context is then defined by comparing (via a Pearson's correlation) its 400-dimensional vector with the average of the vectors of all the preceding words in the corresponding sentence. And the "semantic *dis*similarity" of the word is quantified as 1 minus this correlation (bottom middle left). A vector at the same sampling rate as the recorded neural data is then created that consists of time-aligned impulses at the onset of each word that are scaled according to the value of that word's semantic dissimilarity. The ongoing EEG data are then regressed against this vector to obtain a so-called temporal response function (TRF; right) that describes via beta weights how fluctuations in semantic dissimilarity across words impact upon the EEG at various time lags [21].

was the first word in a sentence, we compared it to the average of all word vectors in the previous sentence and again subtracted the correlation from 1. This produced a single semantic dissimilarity measure for each word that acts as a representation of the meaning added to a sentence by that word (technically, this could take any value between 0 and 2, but it tended to be in the range 0.53–1.06). We then created a vector at the same sampling rate as our EEG data (128 Hz), which consisted of time-aligned impulses at the onset of each word that were scaled according to the value of that word's semantic dissimilarity. Then, by linearly regressing the low-frequency (1–8 Hz) EEG against this vector, we derived a so-called temporal response function (TRF) [20] that describes how these fluctuations in semantic dissimilarity across consecutive words impact upon the neural activity at various time lags (see Figure 1).

A Neural Correlate of Semantic Dissimilarity in Natural Speech

A TRF averaged over 19 subjects who each listened to ~60 min of an audiobook is shown in Figures 2A and 2B. A prominent negativity is apparent over midline parietal scalp at time lags between 200 and 500 ms (Figure 2A). Over this time range, this negativity was significantly less than zero across subjects at several parietal scalp electrode sites (Figure 2B; running one-tailed t test, $p < 0.05$, false discovery rate [FDR] corrected). To confirm that this negativity was related to the semantic content of the speech and not just the stimulus acoustics, we repeated the analysis using another dataset from ten subjects who listened to the same audiobook, but in a time-reversed fashion [21]. Carrying out the same steps as before (while taking into account the time-reversed nature of the stimuli) produced TRF responses that showed no evidence of the prominent, late negativity (Figures 2A and 2B; running one-tailed t test, $p > 0.05$, FDR corrected). The presence and absence of this negativity for forward and time-reversed speech, respectively, was evident for several individual subjects who undertook both experiments (Figure 2C). These results demonstrate that electrophysiological responses to natural speech, in the form of a late, parietal negativity, reflect the semantic dissimilarity of individual words to their preceding context in a relatively tightly time-locked fashion.

The TRF analysis is based on a linear regression between the stimulus feature (i.e., semantic dissimilarity) and the EEG response. We wished to determine how reasonable this linear assumption was or whether, in fact, our TRF negativity was being driven primarily by occasional "incongruent" words with large semantic dissimilarity values. To do this, we divided the semantic dissimilarity impulses into four quartiles based on their magnitude. We then created four separate regressors corresponding to each of these subsets of impulses, where the impulses within each subset were rescaled to have unit height. The amplitude of the late negativity in the TRFs for these four quartiles showed a monotonic increase with measure of semantic dissimilarity (Figure S1).

Neural Signatures of Semantic Dissimilarity Depend on Intelligibility

The experiments above involved either completely intelligible or completely unintelligible stimuli. To assess how sensitive our semantic dissimilarity TRF might be to gradations of intelligibility, we reanalyzed data from another experiment involving speech-in-noise [22]. Specifically, we analyzed EEG data from 21

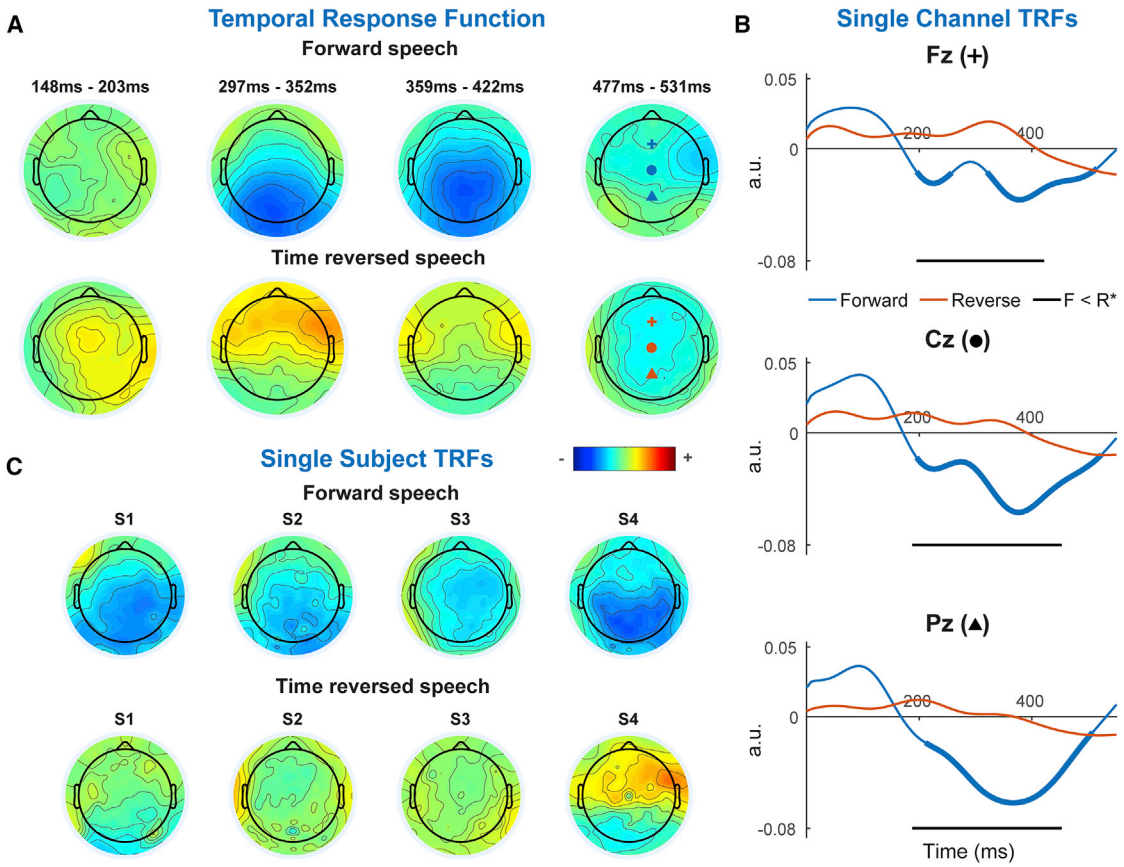

Figure 2. Temporal Response Functions for Natural and Time-Reversed Speech

(A) Topographic maps of the semantic dissimilarity TRF averaged over all trials and all subjects for natural, forward speech (top) display a marked centro-parietal negativity between ~200 and 480 ms. There is no evidence of a similar negativity in the average TRF for time-reversed speech (bottom). Further tests on the assumption of linearity in the TRF model were conducted and shown in Figure S1.

(B) Grand average TRF waveforms at selected individual channels show the time course of the negativity related to semantic dissimilarity. Thick lines indicate a response that is statistically less than zero across subjects ($p < 0.05$, t test, FDR corrected). Black lines below the waveforms indicate that the TRFs for forward speech are statistically more negative than those for time-reversed speech across subjects ($p < 0.05$, t test, FDR corrected).

(C) Topographic maps of TRFs averaged over the interval 200–500 ms for selected subjects who took part in both the forward and time-reversed speech experiments. For all four subjects, a negativity is apparent for forward speech (albeit with slightly different distributions for each subject) that is absent for time-reversed speech. A further study relating this new measure to "the classic N400" is shown in Figure S2.

subjects who listened to two repetitions each of fifteen 60 s segments of continuous audio speech which were always mixed with spectrally matched stationary noise at a signal-to-noise ratio of −9 dB. Based on the well-known fact that visual speech enhances the intelligibility of speech-in-noise [23], we manipulated intelligibility by allowing subjects to watch a video of the speaker for one of these repetitions. And although the audio-only speech was not completely unintelligible, the presence of the video led to a large, significant improvement in intelligibility as measured by both self-report ($p = 4 \times 10^{-5}$, Wilcoxon signed-rank test) and a word-detection task ($p = 7.9 \times 10^{-5}$, Wilcoxon signed-rank test). This behavioral effect was mirrored by a significant difference in the semantic dissimilarity TRFs between audio-only and audiovisual conditions (Figures 3A and 3B). This difference was most pronounced at time lags between 380 and 600 ms, where it showed an effect size of $d' = 0.55$. Notably, this was substantially later than the interval for the TRF negativity during clean speech (Figure 2), a fact that may have to do with the increasing difficulty of processing speech under noisy conditions (see [6, 25]). Another way to quantify how intelligibility affects brain activity is to assess how well our semantic dissimilarity TRFs can predict unseen EEG responses to natural speech. This kind of forward-encoding model-based approach has previously been used for predicting EEG responses based on envelope and phonetic representations of speech [21], as well as fMRI activations based directly on semantic speech vectors [26, 27] and semantic distance [17]. Using cross-validation to fit and test a semantic dissimilarity TRF produced a significantly better EEG prediction for audiovisual speech than audio speech on midline parietal electrode channels (Figure 3C; $p = 0.01$, Wilcoxon signed-rank test). And although the EEG predictions based on audio speech were significantly greater than zero—after all, the audio-alone speech was not completely unintelligible (and see [25])—the effect size of adding the visual input on these EEG predictions scores was large ($d' = 0.84$ on midline parietal electrode Pz). Overall, this demonstrates that our semantic

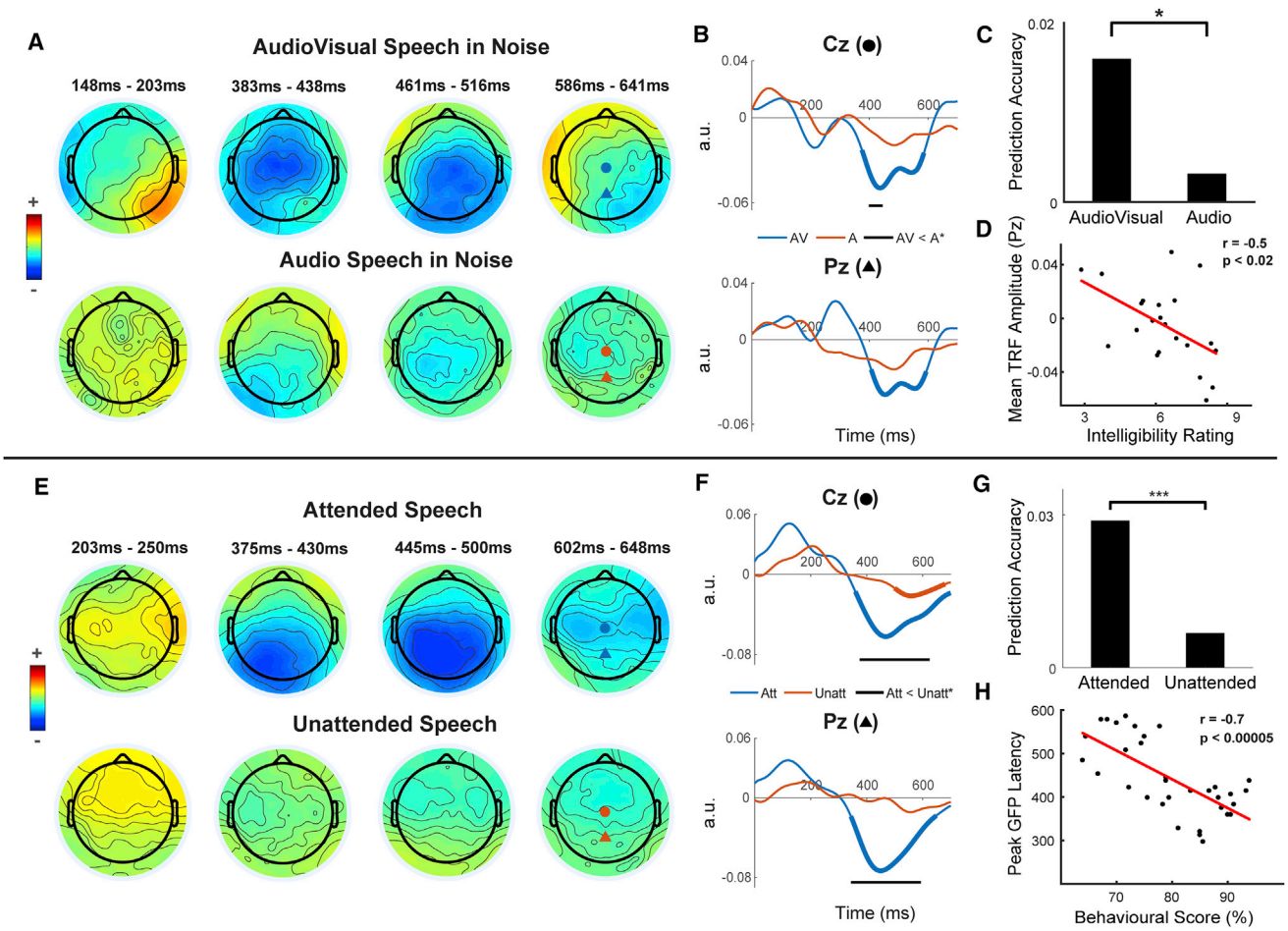

Figure 3. Assessing the Effect of Comprehension on the Electrophysiological Index of Semantic Dissimilarity

(A) Topographic maps of the semantic dissimilarity TRF averaged over all trials and all subjects for audiovisual speech in −9 dB of acoustic background noise display a centro-parietal negativity between ~400 and 600 ms. This negativity is significantly reduced in the average TRF for audio-only speech in the same level of background noise, which was much less intelligible.

(B) Grand-average TRF waveforms for audiovisual and audio-only speech over two selected midline electrodes. Thick lines indicate a response that is statistically less than zero across subjects ($p < 0.05$, running t test, FDR corrected). Black lines below the waveforms indicate that the TRFs for audiovisual speech are statistically more negative than those for audio-only speech across subjects ($p < 0.05$, running t test, FDR corrected).

(C) A cross-validation procedure was used to predict EEG responses to natural speech using a semantic dissimilarity TRF trained on other data. EEG prediction accuracy for audiovisual speech was significantly greater than that for audio-only speech ($p < 0.01$, t test).

(D) Across subjects, the amplitude of the semantic dissimilarity TRF over midline parietal scalp was significantly correlated with self-reported intelligibility rating of audiovisual speech ($p < 0.02$, Pearson's correlation).

(E) Topographic maps of the semantic dissimilarity TRF averaged over all trials and all subjects for attended speech in a dichotic cocktail party paradigm display a centro-parietal negativity between ~300 and 600 ms. This negativity is not apparent in the average TRF for unattended speech.

(F) Grand average TRF waveforms for attended and unattended speech over two selected midline electrodes. Thick lines indicate a response that is statistically less than zero across subjects ($p < 0.05$, running t test, FDR corrected). Black lines below the waveforms indicate that the TRFs for attended speech are statistically more negative than those for unattended speech across subjects ($p < 0.05$, running t test, FDR corrected).

(G) A cross-validation procedure was used to predict EEG responses to natural speech using a semantic dissimilarity TRF trained on other data. EEG prediction accuracy for attended speech was significantly greater than that for unattended speech ($p < 1 \times 10^{-6}$, t test).

(H) Across subjects, the latency of the peak in the global field power (GFP) [24] of the semantic dissimilarity TRF was significantly negatively correlated with the number of questions answered correctly on the attended speech ($p < 5 \times 10^{-5}$, Pearson's correlation).

dissimilarity TRF is sensitive to variations in the intelligibility of acoustically identical speech. Moreover, in the audiovisual speech condition, there was a significant negative correlation across subjects between the self-reported intelligibility ratings (which varied broadly) and the amplitude of the TRF negativity averaged over the interval 250–500 ms (Figure 3D; the more intelligible, the larger the negativity; $r = -0.5$, $p < 0.02$).

No Evidence of Contextual Semantic Processing for Unattended Speech

Over 60 years ago, it was first noted that, when attending to one of two dichotically presented speech streams, people have a very limited ability to report on the content of the speech in the unattended ear [28], a phenomenon known as the cocktail party effect. Ever since then, researchers have sought to explain this

phenomenon in terms of psychological models [29–31] and neurophysiological data [32–35]. Despite these efforts, the extent to which unattended speech is semantically processed by the brain remains unclear [24, 36]. However, given the very marked limitations in the ability of subjects to report on the content of unattended speech, we hypothesized that the negativity in our TRF, as an index of contextual semantic processing, should be markedly reduced in unattended speech.

We reanalyzed EEG recorded from 33 subjects who attended to one of two concurrently and dichotically presented audiobooks (17 subjects attended to one story and 16 to the other) [33, 37]. The experiment was paused after every ~60 s, and subjects were asked multiple-choice questions on both stories. We derived semantic dissimilarity regressors for each of the two stories, and then regressed the EEG data against these vectors to produce two TRFs—one for the attended story and one for the unattended story. Consistent with previous studies, the behavioral effect of attention was very strong (80% correct answers for the attended story and 29% for the unattended story—chance was 25%). Mirroring this large behavioral effect, the TRF corresponding to the attended story showed a clear and prominent negativity over midline parietal scalp that was significantly larger than the corresponding TRF for the unattended speech, again at a rather long latency of 380–600 ms (Figures 3E and 3F; paired t test, $p = 9.3 \times 10^{-8}$; effect size $d' = 2.0$). Although this does not entirely rule out some level of semantic processing in unattended speech—after all, our regressors are based on one particular computational measure of linguistic processing—it does present strong evidence of a pronounced reduction in the processing of unattended words relative to their context. Once more, using cross-validation to fit and test a semantic dissimilarity TRF produced a significantly better EEG prediction for the attended speech than the unattended speech (Figure 3G; $p = 9.36 \times 10^{-7}$, Wilcoxon signed-rank test). And while the EEG predictions based on unattended speech were significantly greater than zero—possibly as a result of weak correlations between the semantic dissimilarity impulses and acoustic energy changes at word onsets—the effect size of attention on these EEG prediction scores was as large as that on the TRFs themselves ($d' = 2.0$ at electrode Pz). Unlike for the audiovisual speech-in-noise experiment above, we found no relationship between the amplitude of the TRF and performance on the attended questions. This was unsurprising given that the to-be-attended speech stream was always intelligible. Instead, we found that the peak latency of the TRF negativity was significantly negatively correlated with performance on the questions across subjects ($r = -0.7$, $p = 1.95 \times 10^{-5}$). In other words, the earlier a subject's TRF peak, the better that subject did on the task. We interpret this as evidence that people who can successfully sustain their attention and/or suppress distracting information can more efficiently process the behaviorally relevant speech—or vice versa. This notion of more efficient semantic processing of words in their recent historical context aligns with the well-known link between working memory and cocktail party attention performance [38].

On the Relationship of the Semantic Dissimilarity TRF Negativity to the N400 Component

The dominant feature of our semantic dissimilarity TRF is a negativity over centro-parietal scalp, a feature shared by the EEG measure that has most strongly been linked with semantic processing—the N400 component. Although this correspondence is not terribly surprising given that the derivation of our semantic dissimilarity measure was motivated by the fact that the N400 is elicited by words that are semantically incongruent with their context, we have refrained from referring to our response as being the classic N400. The reason for this is that the specific assumptions about semantic dissimilarity that underlie our TRF analysis are not precisely the same as the differences in predictability (cloze probability [39]) that drive the development of most N400 stimuli [4]. As such, although it is possible that our approach is simply another way to derive the classic N400, it is also possible that the two measures may reflect at least partially dissociable processes. As a first attempt to examine the relationship between the two responses more directly, we recorded EEGs from nine subjects who undertook a classic N400 experiment and who listened to the audiobook used in our first-mentioned experiment above. For the N400 experiment, subjects read 300 sentences presented word by word on a screen, half of which ended with a word that was congruent with the rest of the sentence and half which ended with an incongruent word. N400s were then determined by subtracting the event-related potential to the congruent words from that to the incongruent words. And, using the EEG data recorded during the story, we derived a semantic dissimilarity TRF for each subject as before. The two responses displayed somewhat similar time courses over midline parietal scalp (although the TRF peaked significantly earlier than the N400; $p = 0.012$; Figure S2), as well as similar topographical distributions at 375–425 ms. The peak amplitude of the N400 component was also significantly correlated with the peak amplitude of the semantic dissimilarity TRF across the nine subjects ($r = 0.751$, $p = 0.02$; Figure S2).

DISCUSSION

We have shown that when listening to natural speech, the ongoing dynamics of cortical activity reflect the semantic processing of words in context in a rapid, time-locked fashion. And we have shown that indices of this processing are robustly affected by whether subjects understand the speech they hear and whether they are paying attention to that speech. This approach adds an extra dimension to research on the neural tracking of natural speech dynamics by directly linking a new component of that tracking to the contextual semantic processing of speech. Further work will be necessary to more fully characterize this online semantic processing. This will include investigating whether other types of language knowledge contribute to our measures [17], assessing whether unattended speech is processed at a semantic level that depends less upon context than our dissimilarity measure, and modeling semantic representation using more neurobiologically motivated approaches [40] rather than just word co-occurrence. By incorporating other computational models into the framework we have outlined, we would expect that EEG, electrocorticography (ECoG), and magnetoencephalography (MEG) could be very useful in answering these questions.

Gaining deeper insights on these issues will also be helped by optimizing our analysis. For example, it is likely that generating

regressors by aligning impulses to word onsets, as we have done, is suboptimal. This is because the time taken to semantically process different words necessarily depends on the words themselves, their context, and listening conditions. That said, the assumption underlying our analysis is that for the successful comprehension of natural speech to be possible at all, meaning must typically be extracted in a relatively online manner—otherwise, we would constantly lose track of what we are hearing. And this assumption appears to be borne out by existence of the robust TRFs we observe. We hope that the framework we have introduced here will allow researchers to estimate more optimal locations for the regressor events and/or regressor events other than impulses. For example, it may be possible to use a "reverse" approach to discover such optimal regressors [41].

It will also be important to more fully examine the relationship between our TRF and the N400. It is entirely possible that they are effectively functionally equivalent. That said, although the amplitude correlations between the two measures were significant, they were decidedly imperfect, and their peak latencies also differed significantly. Now these differences may be due to the simple fact that the TRF was derived using audio speech, whereas the N400 was elicited using reading. But it also remains possible that the differing assumptions used in producing each measure mean they are reflecting different sub-processes involved in semantic understanding. Future work is needed to answer this question. This includes efforts to determine what sub-processes the TRF negativity represents, similar to efforts involving the N400 [42, 43]. Either way, it remains the case that our study has introduced a novel framework for extracting EEG responses that reflect the semantic processing of natural, running speech. This framework could be useful in a broad range of basic, applied, and clinical research studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data Acquisition and Pre-processing
  - Stimuli and Procedures
  - Computational Model and Regression
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and can be found with this article online at https://doi.org/10.1016/j.cub.2018.01.080.

ACKNOWLEDGMENTS

This study was supported by the Irish Research Council through the Government of Ireland Postgraduate Scholarship scheme (GOIPG/2016/1553 to M.P.B. and GOIPG/2013/1249 to G.D.L.) and a Career Development Award from Science Foundation Ireland (15/CDA/3316; E.C.L.).

AUTHOR CONTRIBUTIONS

E.C.L., G.D.L., and A.J.A. conceived of the experiment. M.B, G.D.L., and M.J.C. collected data. M.B., G.D.L., and A.J.A. analyzed the data. E.C.L. wrote the first draft of the manuscript. M.B., G.D.L. and A.J.A edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 9, 2017
Revised: December 18, 2017
Accepted: January 29, 2018
Published: February 22, 2018

REFERENCES

1. Crystal, T.H., and House, A.S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. J. Acoust. Soc. Am. *88*, 101–112.
2. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychol. Rev. *74*, 431–461.
3. Hagoort, P., and Brown, C.M. (2000). ERP effects of listening to speech: semantic ERP effects. Neuropsychologia *38*, 1518–1530.
4. Kutas, M., and Federmeier, K.D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. *62*, 621–647.
5. Friederici, A.D., Pfeifer, E., and Hahne, A. (1993). Event-related brain potentials during natural speech processing: effects of semantic, morphological and syntactic violations. Brain Res. Cogn. Brain Res. *1*, 183–192.
6. Strauß, A., Kotz, S.A., and Obleser, J. (2013). Narrowed expectancies under degraded speech: revisiting the N400. J. Cogn. Neurosci. *25*, 1383–1395.
7. Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M.M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. USA *98*, 13367–13372.
8. Lalor, E.C., and Foxe, J.J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. *31*, 189–193.
9. Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron *54*, 1001–1010.
10. Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. Nature *244*, 522–523.
11. Simpson, G.B. (1991). Understanding Word and Sentence*Volume 77* (Elsevier).
12. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., and Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science *268*, 1632–1634.
13. Landauer, T.K., and Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. *104*, 211.
14. Levy, R. (2008). Expectation-based syntactic comprehension. Cognition *106*, 1126–1177.
15. Pynte, J., New, B., and Kennedy, A. (2008). On-line contextual influences during reading normal text: a multiple-regression analysis. Vision Res. *48*, 2172–2183.
16. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient estimation of word representations in vector space. arXiv, arXiv:1301.3781, http://arxiv.org/abs/1301.3781.
17. Frank, S.L., and Willems, R.M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Lang. Cogn. Neurosci. *32*, 1192–1203.

18. Mitchell, J., and Lapata, M. (2010). Composition in distributional models of semantics. Cogn. Sci. *34*, 1388–1429.
19. Kiela, D., and Clark, S. (2014). A systematic study of semantic vector space model parameters. Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL, pp. 21–30.
20. Crosse, M.J., Di Liberto, G.M., Bednar, A., and Lalor, E.C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front. Hum. Neurosci. *10*, 604.
21. Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. *25*, 2457–2465.
22. Crosse, M.J., Di Liberto, G.M., and Lalor, E.C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J. Neurosci. *36*, 9888–9895.
23. Sumby, W.H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. *26*, 212–215.
24. Aydelott, J., Jamaluddin, Z., and Nixon Pearce, S. (2015). Semantic processing of unattended speech in dichotic listening. J. Acoust. Soc. Am. *138*, 964–975.
25. Ding, N., and Simon, J.Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. *33*, 5728–5735.
26. de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The hierarchical cortical organization of human speech processing. J. Neurosci. *37*, 6539–6557.
27. Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., and Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458.
28. Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. *25*, 975–979.
29. Broadbent, D.E. (1958). Perception and Communication (Pergamon Press).
30. Deutsch, J.A., and Deutsch, D. (1963). Attention: some theoretical considerations. Psychol. Rev. *70*, 80–90.
31. Treisman, A.M. (1964). Verbal cues, language, and meaning in selective attention. Am. J. Psychol. *77*, 206–219.
32. Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature *485*, 233–236.
33. Power, A.J., Foxe, J.J., Forde, E.J., Reilly, R.B., and Lalor, E.C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. Eur. J. Neurosci. *35*, 1497–1503.
34. Teder, W., Kujala, T., and Näätänen, R. (1993). Selection of speech messages in free-field listening. Neuroreport *5*, 307–309.
35. Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron *77*, 980–991.
36. Lachter, J., Forster, K.I., and Ruthruff, E. (2004). Forty-five years after Broadbent (1958): still no identification without attention. Psychol. Rev. *111*, 880–913.
37. O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., and Lalor, E.C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb. Cortex *25*, 1697–1706.
38. Conway, A.R., Cowan, N., and Bunting, M.F. (2001). The cocktail party phenomenon revisited: the importance of working memory capacity. Psychon. Bull. Rev. *8*, 331–335.
39. Taylor, W.L. (1953). "Cloze procedure": a new tool for measuring readability. Journalism Bulletin *30*, 415–433.
40. Anderson, A.J., Binder, J.R., Fernandino, L., Humphries, C.J., Conant, L.L., Aguilar, M., Wang, X., Doko, D., and Raizada, R.D.S. (2017). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cereb. Cortex *27*, 4379–4395.
41. Huth, A.G., Lee, T., Nishimoto, S., Bilenko, N.Y., Vu, A.T., and Gallant, J.L. (2016). Decoding the semantic content of natural movies from human brain activity. Front. Syst. Neurosci. *10*, 81.
42. Lau, E.F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. Nat. Rev. Neurosci. *9*, 920–933.
43. Lau, E.F., Holcomb, P.J., and Kuperberg, G.R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. J. Cogn. Neurosci. *25*, 484–502.
44. Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods *134*, 9–21.
45. Crosse, M.J., Butler, J.S., and Lalor, E.C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. *35*, 14195–14204.
46. Ding, N., Chatterjee, M., and Simon, J.Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage *88*, 41–46.
47. Parsons, T.W. (1987). Voice and Speech Processing (McGraw-Hill College).
48. Baroni, M., Dinu, G., and Kruszewski. G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 1, Long Papers (Association for Computational Linguistics), pp. 238–247.
49. Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-aligner: a tool for forced alignment of laboratory speech. Can. Acoust. *39*, 192–193.
50. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
EEG data, stimuli and behavioral measures	Dryad	https://doi.org/10.5061/dryad.070jc
Software and Algorithms		
MATLAB 2016b	The MathWorks	https://uk.mathworks.com/
Presentation Software	Neurobehavioral Systems	https://www.neurobs.com/
mTRF Toolbox	M.J. Crosse, G.M. Liberto, E.C. Lalor	https://sourceforge.net/projects/aespa/
Other		
Sennheiser HD650 headphones	Sennheiser Electronic	N/A
'The Old Man and the Sea' by Ernest Hemingway	Charles Schribner's Sons	N/A
'20000 Leagues Under the Sea' by Jules Verne	Pierre-Jules Hetzel	N/A
'Journey to the Center of the Earth' by Jules Verne	Pierre-Jules Hetzel	N/A
Obama's Weekly Address	The White House	https://obamawhitehouse.archives.gov/briefing-room/weekly-address

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Michael Broderick (brodermi@tcd.ie). There is no restriction for distribution of materials.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All subjects were native English speakers, and reported normal hearing, normal or corrected-to-normal vision, and no history of neurological disease. 19 subjects (13 male) aged between 19 and 38 years participated in the first experiment involving listening to a single audiobook. Of these 19 subjects, 9 also participated in the N400 experiment. Ten subjects (7 male) aged between 21 and 32 years participated in the experiment involving the time-reversed audiobook (5 of these subjects participated in the first experiment above). 34 subjects (28 male) with a mean age of 27.3 ± 3.2 (SD) years participated in the cocktail party attention experiment, but data from one subject was not included in the analysis as the recordings from their mastoid electrodes were of poor quality. 21 subjects (6 female) aged between 21 and 35 years participated in the multisensory speech experiment. Much of the data from these experiments has previously been published in studies examining how EEG tracks the envelope and phonetic content of speech [21, 22, 32, 37]. All procedures were undertaken in accordance with the Declaration of Helsinki and were approved by the Ethics Committees of the School of Psychology at Trinity College Dublin, and the Health Sciences Faculty at Trinity College Dublin.

METHOD DETAILS

Data Acquisition and Pre-processing

For all experiments, 128-channel EEG data (plus two mastoid channels) were acquired at a rate of 512 Hz using an ActiveTwo system (BioSemi). Triggers indicating the start of each trial were sent by the stimulus presentation computer and included in the EEG recordings to ensure synchronization. Offline, the data were band-pass filtered between 1 and 8 Hz, downsampled to 128 Hz, and re-referenced to the average of the mastoid channels in MATLAB. To identify channels with excessive noise, the time series were visually inspected and the SD of each channel was compared with that of the surrounding channels. Channels contaminated by noise were recalculated by spline interpolating the surrounding clean channels in EEGLAB [44].

Stimuli and Procedures

In the first experiment, subjects undertook 20 trials, each of the same length (just under 180 s), where they were presented with a professional audio-book version of a popular mid-20th century American work of fiction written in an economical and understated style and read by a single male American speaker. The trials preserved the storyline, with neither repetitions nor discontinuities. The average speech rate was ~210 words/min. Similarly, the second experiment involved the presentation of the same trials in the same order, but with each of the 28 speech segments played in reverse. All stimuli were presented monophonically at a sampling rate of 44.1 kHz using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems

(http://www.neurobs.com). Testing was carried out in a dark, sound-attenuated room and subjects were instructed to maintain visual fixation on a crosshair centered on the screen for the duration of each trial, and to minimize eye blinking and all other motor activities. Data from 10 of the subjects (aged 23–38 years; 7 male) who participated in the first experiment and all of the subjects (aged 21–32 years; 7 male) who participated in the second experiment have been published previously [21]. Data from an additional 9 subjects (aged 19–32 years, 6 male) for the first experiment were collected for the current study.

In the N400 experiment, subjects read 300 sentences presented word-by-word on a screen, half of which ended with a word that was congruent (high cloze probability) with the rest of the sentence and half which ended with an incongruent (low cloze probability) word. N400s were then determined by subtracting the event-related potential to the congruent words from that to the incongruent words. And, using the EEG data recorded during the story, we derived a semantic dissimilarity TRF for each subject as before.

In the cocktail party experiment, 33 subjects (aged 23–38 years; 27 male) undertook 30 trials, each of 60 s in length, where they were presented with 2 classic works of fiction: one to the left ear, and the other to the right ear. Each story was read by a different male speaker. Subjects were divided into 2 groups of 17 and 16 (+1 excluded subject) with each group instructed to attend to the story in either the left or right ear throughout the entire 30 trials. After each trial, subjects were required to answer between 4 and 6 multiple-choice questions on both stories. Each question had 4 possible answers. We used a between-subjects design as we wanted each subject to follow just one story to make the experiment as natural as possible and because we wished to avoid any repeated presentation of stimuli. For both stories, each trial began where the story ended on the previous trial. Stimulus amplitudes in each audio stream within each trial were normalized to have the same root mean squared (RMS) intensity. In order to minimize the possibility of the unattended stream capturing the subjects' attention during silent periods in the attended stream, silent gaps exceeding 0.5 s were truncated to 0.5 s in duration. Stimuli were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com). Subjects were instructed to maintain visual fixation on a crosshair centered on the screen for the duration of each trial, and to minimize eye blinking and all other motor activities. The data from all subjects in this experiment have been published previously [33, 37].

For the multisensory experiment, the stimuli were drawn from a set of videos that consisted of a male speaking American English in a conversational-like manner. Fifteen 60 s videos were rendered into 1280 × 720-pixel movies at 30 frames/s and exported in audio-only (A), visual-only (V), and AV format in VideoPad Video Editor (NCH Software). The soundtracks were sampled at 48 kHz, underwent dynamic range compression, and were matched in RMS intensity (see [45]), and were mixed with spectrally-matched stationary noise to ensure consistent masking across stimuli [25, 46]. The noise stimuli were generated in MATLAB (The MathWorks) using a 50th-order forward linear predictive model estimated from the original speech recording. Prediction order was calculated based on the sampling rate of the soundtracks [47]. The data analyzed here were from the A and AV condition only. Please note, the presentation order of A, and AV repetitions was randomized across the 15 videos and across subjects. The data from all 21 subjects (aged 21–35 years; 13 male) in this experiment have been published previously [22].

Computational Model and Regression

Semantic vectors for content words were derived using the state-of-the-art word2vec algorithm [16]. The ''continuous bag of words'' implementation built in [48] was selected because this was trained on British English corpora (ukWaC, the English Wikipedia and the British National Corpus combined) which is both large and probably more reflective of the language exposure of the participants (in Dublin) than US corpora. In addition, word vectors are freely downloadable (see [48]). Word2vec embodies the ''distributional hypothesis'' that words with similar meaning occur in similar contexts in an artificial neural network approach. Practically, the approach involves sliding a fixed window of words (11 in this case, however this is a parameter set by the experimenter) over a text corpus and training a neural network to predict the word in the center of that window. Word identity (as opposed to semantics) is uniquely encoded as a single bit set to one in a long vector of zeros (vector length is the number of words in the vocabulary). These long vectors form the basis of the input and output to the neural network. The input corresponds to the sum of the 10 word vectors in the window, the output is the central word. Because word order is lost in this summation, the input is analogous to an unordered bag of words. The network contains an internal hidden layer of 400 dimensions. The hidden layer is fully connected to the input and output. It is in fact the weights on the connections between the input and hidden layer that are ultimately harvested to form the semantic model (the weights are a number-of-words in the vocabulary by 400 floating point matrix) and the remainder of the network is discarded. Weights are initially set as random, but are subsequently optimized so as to reduce error between predicted and target output. Intuitively, because words that frequently appear together in the same context window also predict similar central words, weights on these words are tuned to similar internal representations reflecting common contexts. For more details on the training procedure see [48] and [16] (note, the choice of 400 dimensions for the internal layer was arbitrary and, as described in the next paragraph, these 400-dimensional vectors are reduced to a single correlation measure).

Having obtained a vector for each word, we then quantified how semantically dissimilar each particular word was to the preceding words in the corresponding sentence. We did this by calculating a Pearson's correlation between the word's 400-dimensional vector and the average of the vectors corresponding to all the preceding words in that particular sentence, and subtracting this correlation from 1 (where a specific word was the first word in a sentence, we calculated the correlation between the word and the average of all word vectors in the previous sentence, before, again, subtracting that correlation from 1). It should be noted that, this kind of simple feature-wise averaging/summation of word-level semantic vectors has proven to be an effective and enduring method of modeling semantic composition in computational linguistics (e.g., [18, 19]). It has also been proven to be a successful method for predicting fMRI activation patterns associated with sentences' meanings (e.g., [40]). However, it should also be noted that the approach is a

gross oversimplification of the complexities of semantic composition in the brain (and does not take into account the effects of word order or syntax – see discussion [40]). In any case, our approach produced a single semantic dissimilarity measure for each word with a value between 0 and 2. We then created a ''semantic dissimilarity vector'' at the same sampling rate as our EEG data (128 Hz) which consisted of time-aligned impulses at the onset of each word that were scaled according to the value of that word's semantic dissimilarity. The word onset times were determined by performing forced alignment of the speech files and the corresponding textual orthographical transcription using the Prosodylab-Aligner, which has been shown to produce alignments with median precision (misalignment) on the order of 10 ms [49].

A system identification technique was used to compute a channel-specific mapping between the semantic dissimilarity vector and the recorded EEG data, commonly referred to as a temporal response function (TRF). A TRF can be interpreted as a filter that describes the brain's linear transformation of a stimulus feature, $S(t)$, to the continuous neural response, $R(t)$, over a series of specified time lags: $R(t) = TRF^{*}S(t)$, where '*' represents the convolution operator. Specifically, estimation of the TRF weights was performed using regularized linear regression, wherein a regularization (ridge) parameter was tuned to control overfitting (see [20] for a detailed description of this step).

In previous work, we have attempted to cast our TRF functions with μV as their unit of measure. However, this relies on a decision to normalize the input stimulus values between some limits and, as such, has been somewhat arbitrary. In the present work, and in line with previous work from other groups, the EEG data on each channel was z-scored prior to estimating the TRF, meaning that the TRFs are ultimately presented in arbitrary units. The colors in the TRF topographic plots can be interpreted as follows: red at a particular latency indicates that, at that poststimulus lag, the EEG voltage is driven in a positive direction by a unit change in semantic dissimilarity; blue means the EEG voltage at that poststimulus lag is driven negative by a similar change. Thus, given the same normalization strategy for the various speech stimuli used in this study, the TRF responses can be compared in terms of their amplitudes, despite their description in terms of arbitrary units.

QUANTIFICATION AND STATISTICAL ANALYSIS

TRF waveforms in Figures 2B, 3B, 3F and S2A were tested as being significantly less than zero using a running one-tailed t test across subjects. The resulting p values were corrected using the False Discovery Rate (FDR) method [50]. Thicker lines indicate time points of responses that are statistically less than zero ($p < 0.05$). Statistical differences between TRF waveforms for forward versus time-reversed speech (Figure 2B), audio-only versus audiovisual conditions (Figure 3B) and attended versus unattended (Figure 3F) were tested using one-tailed t tests. The resulting p values were FDR corrected. In experiment 2, to evaluate the improvement in intelligibility from audio to audiovisual conditions via self-report and word detection tasks, we used Wilcoxon signed-rank tests. To quantify effect size (*d'*) in comparing mean TRF responses for audio versus audiovisual speech (experiment 2) and attended versus unattended speech (experiment 3) we used Cohen's d (for experiment 2, $n = 21$ and for experiment 3, $n = 33$). The same measure of effect size was used to compare EEG prediction accuracies in experiments 2 (Figure 3C) and 3 (Figure 3G). The difference in magnitude of the negative component between the attended and unattended TRFs was supported by statistical testing across subjects using a paired t test. Correlations between behavioral measures and mean TRF amplitude (Figure 3D) and latency (Figure 3H) were quantified using Pearson's correlation ($n = 21$ and $n = 33$, respectively). Correlations between TRF amplitude and N400 amplitude (Figure S2B) were quantified using Pearson's correlation ($n = 9$). Differences in N400 and TRF peak latency distributions (Figure S2D) were tested using a Wilcoxon signed-rank test.

DATA AND SOFTWARE AVAILABILITY

Our data is available to download via Dryad at https://doi.org/10.5061/dryad.070jc. The TRF analysis was carried out using the freely available multivariate temporal response function (mTRF) toolbox, which can be downloaded from https://sourceforge.net/projects/aespa/.