Reviewed Preprint
v1 • April 17, 2025
Not revised

Reviewed Preprint
v2 • March 11, 2026
Revised by authors

For correspondence:
ines.schoenmann@gmail.com

Competing interests: No competing interests declared

Funding: See page 22

Reviewing editor: Nai Ding, Zhejiang University, China

Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding in naturalistic listening designs

Inés Schönmann[1], Jakub Szewczyk[1], Floris P de Lange[1], Micha Heilbron[1,2]

[1]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands • [2]Amsterdam Brain and Cognition, University of Amsterdam, Amsterdam, Netherlands

eLife Assessment

This **fundamental** study investigates whether neural prediction of words can be measured through pre-activation of neural network word representations in the brain; **convincing** evidence is provided that neural network representations of neighboring words are correlated in natural language. This study urges future studies to carefully differentiate between neural activity that predicts the upcoming word and neural activity that encodes the current words, which contain information that can be used to predict the upcoming word. The study is of potential interest to researchers investigating language encoding in the brain or in large language models.

https://doi.org/10.7554/eLife.106543.2.sa4

Abstract

The human brain is thought to constantly predict future words during language processing. Recently, a new approach emerged that aims to capture neural prediction directly by using vector representations of words (embeddings) to predict brain activity prior to word onset. Two findings have been proposed as hallmarks of neural next-word prediction: (i) significant encoding prior to word onset and (ii) its modulation by word predictability. However, natural language is rife with temporal correlations, where upcoming words share statistical information with preceding ones. This raises a critical question: do these these hallmarks emerge from the brain actively predicting future content, or might they be equally well explained by the regression model exploiting these inherent stimulus dependencies? To distinguish between these alternatives, we applied the same encoding analysis to passive control systems, i.e., representational systems that encode the stimulus but cannot predict upcoming words. We show that both hallmarks emerge in two such control systems, namely in word embeddings themselves and in speech acoustics. We further show that proposed methods to correct for these dependencies are insufficient, as the effects persist even after such corrections. Together, these results suggests that pre-onset prediction of brain activity might reflect dependencies in natural language rather than predictive computations. This questions the extent to which this new encoding-based method can be used to study prediction in the brain.

Introduction

In the past years, the field of natural language processing (NLP) has made great advances in developing computational systems that can generate, classify and interpret language. Much of this progress has been driven by large language models (LLMs): neural networks trained in a self-supervised manner to predict the next word or token (Minaee et al., 2024). Surprisingly, this simple training objective is sufficient for models to learn about language more broadly, making

models develop a human-like knowledge of syntax (Manning et al., 2020; Linzen and Baroni, 2021) and enabling them to solve almost any NLP task (Minaee et al., 2024; Manning, 2022). Further-more, "embeddings", also constitute the highest performing encoding models for predicting brain responses to linguistic stimuli (Caucheteux and King, 2022; Jain and Huth, 2018; Schrimpf et al., 2021)—indicating that LLMs' internal representations might capture some aspects of human language representations (Tuckute et al., 2024).

Two lines of research suggest that predicting the upcoming linguistic input also plays an important role in *human* language comprehension. The first line of research focuses on neural and behavioural responses occurring *after* the onset of the stimulus in question. The reasoning here is that if the brain is engaged in predicting upcoming linguistic input, brain responses and reading times should vary as a function of linguistic predictability (Smith and Levy, 2013; Frank et al., 2015; Willems et al., 2016; Shain et al., 2024; Heilbron et al., 2023). Many studies following this line of reasoning have demonstrated that both neural responses and reading times are sensitive to even subtle fluctuations in predictability at several levels of linguistic analysis, e.g. phonemes, words or semantics (Boston et al., 2011; Brodbeck et al., 2022; Heilbron et al., 2022; Smith and Levy, 2013; Szewczyk and Federmeier, 2022). This approach can be seen as an extension of a long-standing tradition of research into linguistic expectations which relied on carefully constructed sentences that violate linguistic expectations and use cloze probabilities to quantify unexpectedness (Kutas and Hillyard, 1980a,b, 1984).

Recently, a new, alternative approach emerged that aims to probe linguistic predictions directly by capturing their neural signature *prior* to the onset of a word (Wang et al., 2018; Goldstein et al., 2022b). Predicting a word is thought to involve pre-activating the representation of that word. Hence, finding a trace of a representation of a word in the neural signal *prior* to its onset is interpreted as direct evidence for a word's pre-activation, and therefore, next-word prediction. Capturing the neural signature of the prediction itself is appealing, as it has the potential to circumvent interpretational challenges of more indirect, post-stimulus predictability effects which have been suggested to reflect related but distinct downstream processes—such as semantic integration difficulty or 'post-diction'—rather than prediction *per se* (Pickering and Gambi, 2018; Huettig, 2015; Huettig and Mani, 2016). Indeed, one of the most widely used post-stimulus measures of prediction, surprisal, was originally proposed as a measure of syntactic integration difficulty and not as a measure of prediction (Hale, 2001). Perhaps the most influential demonstration of predictive pre-activation during language comprehension was presented by Goldstein et al. (2022b). Using encoding models on electrocorticographic (ECoG) recordings of participants listening to naturalistic speech, they reported two findings which we will refer to as *hallmarks of prediction*: (i) brain responses could be predicted significantly better than chance as early as two seconds prior to word onset, and (ii) this pre-onset encoding was modulated by word predictability, with highly predictable words showing stronger pre-onset encoding. This modulation by predictability is in line with the idea that their representation was pre-activated more strongly. More recently, a similar pattern of results was found in non-invasive MEG recordings (Azizpour et al., 2024).

However, interpreting pre-onset encoding as evidence for prediction is not as straightforward as it may seem. This is because language is rife with temporal dependencies which can potentially be learned by a regression model: Neighbouring words often share semantic content—as in "pine tree" or "green leaves"—or morphosyntactic features, as in "he goes" where both words carry syntactic markers denoting third person singular. Other dependencies are not structural but incidental, such as those caused by words that happen to co-occur together often, such as "Sherlock Holmes". Irrespective of the exact nature of these dependencies, they allow to predict (at least in part) earlier words from subsequent ones. Therefore, they might also allow to predict (at least in principle) brain responses to earlier words using word representations of subsequent words. A priori, then, these inherent stimulus dependencies might already explain why representations of future words can be used to model prior brain responses, without having to assume any predictive pre-activation by the brain. This creates a fundamental ambiguity: Is pre-

onset encoding a reflection of the brain generating predictions about upcoming words, or does it merely show that the regression model can successfully exploit temporal dependencies present in the stimulus material?

One way to distinguish between these alternatives is to use what we call a *passive control system*: a representational system in which the stimulus—and thus its dependencies—are encoded but that, by definition, cannot generate predictions about upcoming words. Speech acoustics provide such an example, as the auditory stimulus is encoded in the speech acoustics, yet they cannot actively 'predict' upcoming words. If, when applying the same analysis to such a control system, the hallmarks of prediction still emerge, this would demonstrate that they must arise from stimulus dependencies alone, without requiring any underlying predictive process.

Here, we directly address this issue. First, we replicate the results reported by Goldstein et al. (2022b) and Azizpour et al. (2024) across two magnetoencaphalography (MEG) datasets, demonstrating that both hallmarks robustly generalise to MEG. We then evaluate two passive control systems—word vectors and speech acoustics—neither of which actively predicts upcoming words. We show that both purported hallmarks emerge in these control systems, despite the absence of any predictive process. Furthermore, we demonstrate that methods proposed to correct for stimulus dependencies, such as removing reoccurring bigrams or residualising neighbouring word information from embeddings, prove insufficient: The proposed hallmarks persist in the acoustic control system even after such corrections. We conclude that both proposed hallmarks can be fully explained by the correlational structure inherent in naturalistic language, without assuming predictive pre-activation in the neural data. This poses a challenge to the use of encoding models for probing linguistic predictions, and questions to what extent such analyses can demonstrate that brains, like LLMs, perform next-word prediction.

Results

Hallmarks of prediction replicate in MEG data

Drawing on two different, publicly available MEG datasets in which participants listened to narratives, we analysed the data following the approach put forth by Goldstein et al. (2022b) and Azizpour et al. (2024). Hence, MEG data were epoched with respect to the onset of each word between −2*s* and +2*s*, and brain activity was averaged over a sliding window of 100*ms*. Subsequently, the word representation (word embedding) corresponding to the word to which each epoch was time-locked, was used to predict brain activity within that epoch, i.e., within the time window of −2*s* to +2*s*(see Figure 1a). Correlating the predicted with the actual brain response results in a time-resolved prediction accuracy curve which allows us to test the two hallmarks of prediction proposed by Goldstein et al. (2022b), namely i) whether brain activity prior to the onset of a word can be predicted from that word's embedding and ii) whether pre-onset prediction accuracy is higher for predictable words than for unpredictable ones. These two aspects are proposed as evidence of prediction, since pre-onset encoding is thought to capture the pre-activation of a word's neural representation.

Using representations from three different models (GPT-2, GloVe and arbitrary 300-dimensional word embeddings) to encode brain activity, we found that it is possible to replicate both hallmarks of predictions described by Goldstein et al. (2022b). We found both i) positive encoding prior to word onset for all three models (see Figure 1B) and ii) a slight encoding advantage prior to word onset for highly predictable (i.e., GPT-2's top-one prediction) as opposed to less predictable words (see Figure 1C). For subject 1, this advantage started as early as 575*ms* prior to word onset, and predictable words led to a continuously higher encoding performance (M = .004, SD = .001, p < 0.031), corresponding to an average improvement in encoding performance within this preonset time window of 17% with respect to unpredictable words (with improvements of M = .006 (SD = .002, p < 0.035) corresponding to 25% starting 525*ms* prior to onset, and M = .011 (SD= .006, p < 0.033) corresponding to 46% during the time window of 1 400 − 300*ms* prior to word onset for subject 2 and 3, respectively, see Figure S1A-B). In accordance with previous findings, using GPT-2's contextualised word representations allowed for earlier and better brain encoding than using

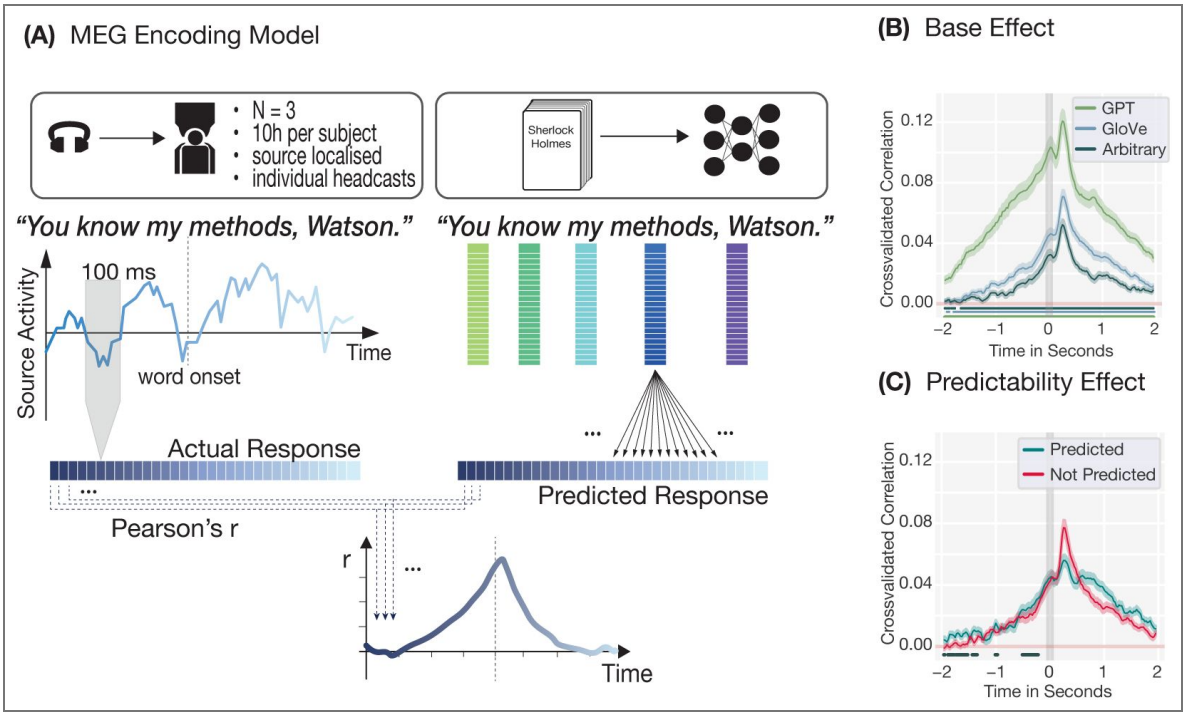


Figure 1.

A) MEG Encoding Model. MEG data was epoched to word onset and averaged over a sliding window of 100*ms*, moving with a step size of 25*ms*. The model representation (GPT-2, GloVe or arbitrary) of the word at $t = 0$ was then used to predict the neural response for each channel and time point in a separate cross-validated Ridge regression. The actual and predicted responses were then correlated time point by time point, resulting in a time-resolved encoding plot. B) Positive pre-onset encoding (subject 1) for GPT-2 (green), GloVe (blue) and arbitrary (grey) embeddings shows that it possible to find ostensible neural signatures of pre-activation in MEG data. Lines show clusters of time points that are significantly different from zero ($p < 0.05$ under the permutation distribution). C) Encoding using GloVe embeddings demonstrate a slight advantage of the predictability of a word (top-one prediction by GPT-2-XL) for pre-onset encoding. Line indicates clusters of time points prior to word onset during which predictable words are significantly better encoded ($p < 0.05$ under the permutation distribution).

non-contextualised GloVe embeddings (Goldstein et al., 2022b; Schrimpf et al., 2020). In turn, these non-contextualized word embeddings predicted brain response better than arbitrary, 300-dimensional vectors. However, arbitrary embeddings still performed remarkably well, given that they contain no structured information besides the identity of a word.

These results replicated for all three participants (see Figure S1A and S1B) in the few-subject dataset, indicating the robustness of the results given sufficient amounts of data (see Figure S1C for results from a more conventional multi-subject dataset, and see discussion section for possible explanations for this discrepancy). This demonstrates that pre-onset encoding is a tremendously robust phenomenon, replicating across different word embeddings (GPT-2, GloVe and even arbitrary embeddings), datasets (single-subject as well as multi-subject) and even different forms of MEG spaces (source as well as sensor data). This suggests that pre-onset encoding accuracy is driven neither by the specific neural data nor by the specific word representations used in the encoding model, further raising the question to which extent the stimulus itself might be driving the effect.

Both hallmarks emerge in passive control systems

Having established that both purported hallmarks generalise to MEG data, we next ask to what extent they can unequivocally be interpreted as reflecting neural pre-activation. For this question, we turn to the control systems: In order for these hallmarks to serve as evidence for signatures of neural predictive processes, they would have to be unique to brain responses. If, however, they can arise from stimulus dependencies alone, they should also appear when applying the same analysis to control systems, i.e. systems in which the stimulus is encoded but which cannot generate predictions.

The first control system we considered, consisted of the word embeddings themselves, namely vector representations of the current word in which linguistic information is encoded but which do not perform any predictive computation. We applied the identical encoding analysis used for modeling the MEG data, but replaced the neural signal with word embeddings at each time point (Figure 2A). The embedding at $t = 0$ was used to predict embeddings at earlier time points. The degree to which an encoding model can predict earlier embeddings from later ones—which we call *self-predictability*—reflects temporal dependencies in the word vector space itself. We tested three types of embeddings: GPT-2 representations, static GloVe vectors, and arbitrary vectors containing no linguistic structure beyond word identity. Note that GPT-2 embeddings are not actually a passive control since these embeddings are the internal representations of a model that is actively predicting the next word. The critical test in this analysis is whether the hallmarks emerge for static and arbitrary embeddings. Nevertheless, we include GPT-2 embeddings for completeness and for comparability to the neural results. We observe that our first control system exhibits both hallmarks of prediction (Figure 2B-C): Pre-onset encoding was significantly above chance for all three vector spaces, and this effect was modulated by word predictability (for a replication of these results for the multi-subject dataset and the stimuli used by Goldstein et al. (2022b) see Figure S2). In other words, preceding word vectors were predicted better when the subsequent word was highly predictable in context—mirroring the pattern observed in neural data, but emerging from stimulus structure alone.

Importantly, however, our self-predictability analysis did not require a mapping between different representational spaces, as is the case when encoding neural data, where word embeddings are used to predict brain responses. Hence, we next tested whether the same hallmarks emerge when using word embeddings to predict another meaningful, passive control system, namely the speech acoustics. Like in the case of word embeddings, the stimulus is encoded in the acoustics, yet the acoustics themselves do not perform any active prediction of upcoming words. Furthermore, since participants listened to the narratives, a representation of the speech acoustics must be present in the neural representation. Consequently, the speech acoustics constitute a meaningful, passive control system when testing for the influence of stimulus dependencies on encoding results.

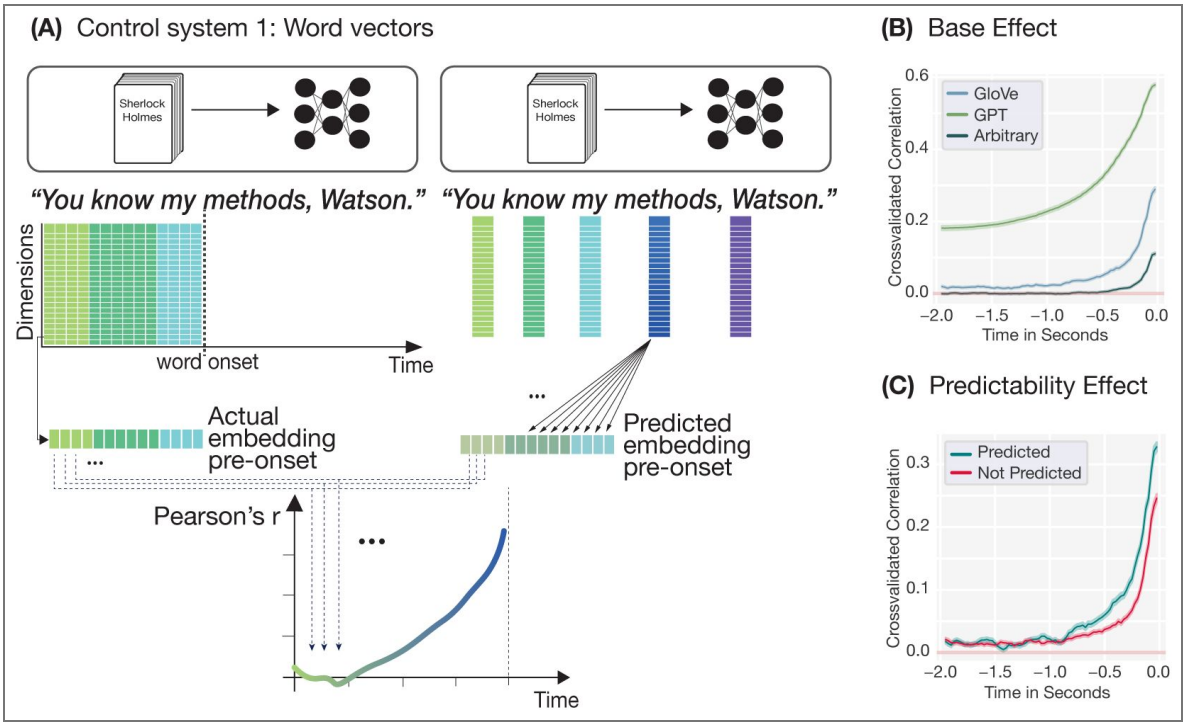


Figure 2.

A) Control system 1: word embeddings. For the first control system, we performed the same analysis as in Figure 1 but replaced the neural data at each time point with a vector representation (embedding) of the word presented at that time point. The word vector at $t = 0$ was then used to predict the previous word vector for each dimension and time point in a separate cross-validated Ridge regression. The actual and predicted values were then correlated time point by time point, resulting in a time-resolved self-predictability plot. B) Pre-onset encoding (*self-predictability*) for GPT-2 (green), GloVe (blue) and arbitrary (grey) embeddings. C) Modulation of pre-onset encoding of static (GloVe) word embeddings by contextual predictability: Prior word vectors are better predicted by successive word vectors if the subsequent word is highly predictable in context (i.e. top-one prediction by GPT-2).

We extracted acoustic features (an 8-band Mel spectrogram and envelope) for each word and applied the same encoding analysis, using the GPT-2, GloVe or arbitrary embedding at $t = 0$ to predict acoustic features at earlier time points (Figure 3A). Again, both hallmarks of prediction emerged in naturalistic speech acoustics (Figure 3B-C, and for a replication of these results for the data used by Goldstein et al. (2022b) see Figure S3B.). Acoustic features prior to word onset could be predicted from the embedding of the upcoming word in all three datasets (Figure 3B, S3), and this effect was modulated by word predictability. In other words, pre-onset acoustics could be predicted better when the subsequent word was highly predictable. Note however, that this modulation did not occur for the acoustic data of our multi-subject dataset, the speech of which was less naturalistic, and for which we could compute only poorer acoustic word representations (see Methods / Discussion).

These findings demonstrate that the proposed hallmarks can emerge in control systems that, by definition, cannot predict upcoming words. Observing these hallmarks in neural data, therefore, does not, by itself, demonstrate that the brain is generating predictions: Hallmarks could instead equally reflect the regression model capitalising on stimulus dependencies which are passively encoded by the brain.

Proposed controls do not effectively remove stimulus dependencies

The previous analyses show that both hallmarks of prediction emerge in passive control systems, indicating that stimulus dependencies alone—rather than next-word prediction—may explain effects found in the neural data. This quite naturally raises the questions whether such stimulus dependencies can be controlled for. A first control already proposed by Goldstein et al. (2022b) was to test whether pre-onset encoding persists even when removing all trials containing reoccurring bigrams for arbitrary embeddings—i.e., random vectors containing no linguistic structure that are assigned to each word. Here, the logic is that nothing but word identity in encoded in arbitrary vectors. Hence, since neighbouring words share no systematic information beyond their incidental co-occurrence in a specific text, removing repeated bigrams eliminates this association. The authors argued that if pre-onset encoding persisted under such conditions, results should be driven by neural pre-activation rather than stimulus dependencies. However, when applying this same constrained analysis, not only did neural encoding remain above chance (see Figure S4A), but crucially, removing reoccurring bigrams did not influence pre-onset encoding in either of our two control systems (see Figure S4B-C). Specifically, across all three datasets, we observe the same results that have hitherto been presented as evidence against encoding stimulus dependencies in our acoustic control system (Figure S5): Arbitrary embeddings could predict pre-onset acoustics significantly above chance, hence encoding the stimulus even after bigram removal. These results demonstrate that this proposed control is insufficient. Removing reoccurring bigrams does not prevent the regression model from encoding stimulus dependencies. Instead our results indicate that regression models can and do exploit dependencies that go beyond mere re-occurrences of bigrams to predict pre-onset features.

A more rigorous approach to removing stimulus dependencies is residualisation: regressing out neighbouring words from each embedding before using it as a predictor in the encoding analysis. For instance, in the sentence “You know my methods”, “You” can be regressed out of “know”, “know” out of “my”, and “my” out of “methods” (see Figure 4A for a visualisation). We residualised each word embedding—thereby removing all information from that embedding that could be predicted linearly from its predecessor. Note that this is a generalised version of a control analysis performed by Goldstein et al. (2022b) which removed neighbouring words through projection. When we recomputed self-predictability using these residualised embeddings, temporal dependencies were successfully eliminated: The encoding model could no longer predict earlier embeddings from later ones (Figure 4B). This confirms that residualisation effectively removes dependencies within the embedding space.

Figure 3.

A) For the second control system we again perform the same analysis as in Figure 1 but replaced the neural data with the stimulus acoustics at that time point. The word embedding of the word at $t = 0$ was then used to predict the prior acoustics. The actual and predicted values were then correlated time point by time point, resulting in a time-resolved correlation plot. B) Pre-onset encoding of speech acoustics based on GPT-2(green), GloVe (blue) and arbitrary (grey) embeddings. C) Modualtion of pre-onset encoding of speech acoustics based on GloVe embeddings by word predictability (top-one prediction by GPT-2-XL).

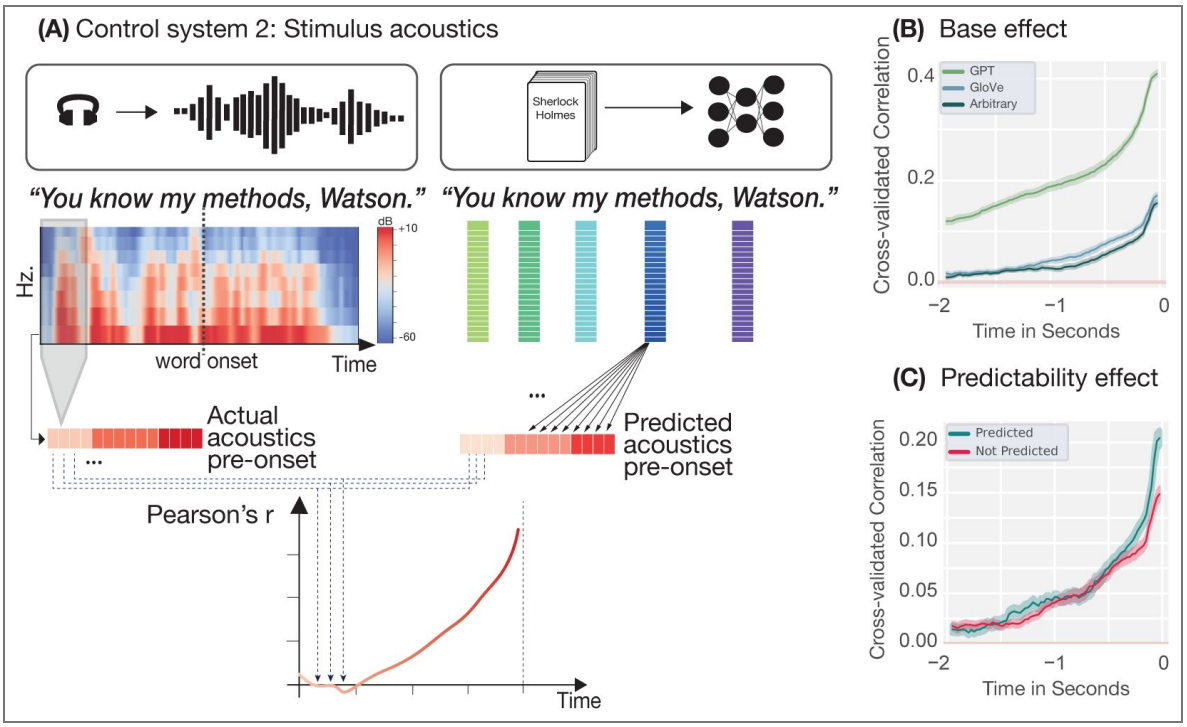
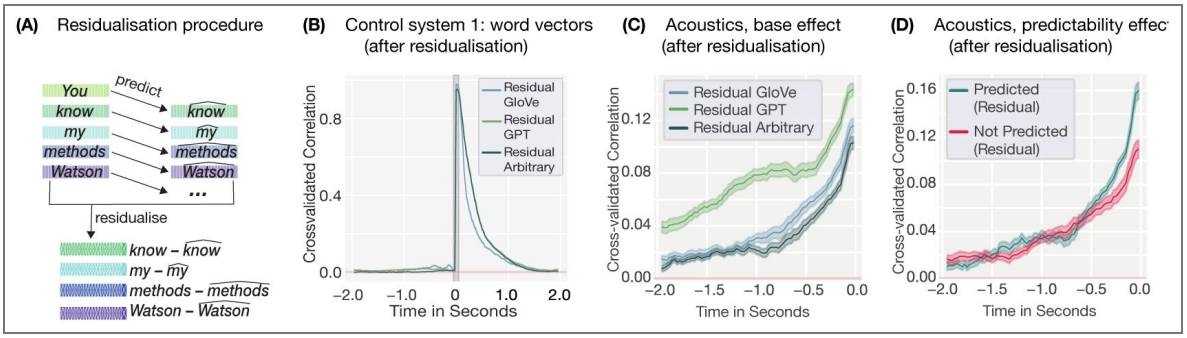


Figure 4.

A) In order to remove shared information between a word and its predecessor in the text, we residualised word embeddings by first fitting an OLS regression to predict the next word based on the previous word's embedding, i.e. predicting "know" based on "You". This resulted in a predicted embedding $\hat{\boldsymbol{x}}$, e.g. "$\widehat{know}$", which contained the shared information between the two words. Finally, the predicted embedding, e.g. "$\widehat{know}$", was removed from the original embedding, e.g. "know", to generate word representations for which the dependency between neighbouring words was removed. B) Self-predictability after regressing out the previous embedding from the embedding at $t = 0$ shows that it is possible to successfully remove the correlations between neighbouring model representations. For brain encoding results when using these residualisded embeddings, see Figure S7. C) Predictability of prior word acoustics when using residualised GPT-2 (green), GloVe (blue) and arbitrary (grey) embeddings prior to word onset. Patterns closely mirror those observed in each model's self-predictability or in residual brain encoding results. D) Predictability of prior word acoustics using residualised GloVe embeddings demonstrate a clear advantage of the predictability of a word (top-one prediction by GPT-2-XL) for predicting its prior acoustic representations, and therefore, the same qualitative difference as observed when encoding neural data.



Strikingly, however, when we used these residualised embeddings to predict our acoustic control system, both hallmarks persisted, insofar as they were present in the original modeling effort (Figure 4C-D and Figure S6). Pre-onset acoustics could still be predicted above chance, and this effect was still modulated by word predictability in both datasets that used natural speech, despite information about neighbouring embeddings being linearly removed. This reveals a fundamental limitation of the approach: Residualisation cannot account for stimulus dependencies when mapping *between* representational spaces. It only removes a very specific representation of neighbouring words from the embeddings, but it fails to remove the underlying dependencies that exist in language itself. However, these underlying dependencies are still present in the system that is modeled, i.e. in the acoustics or the neural data. As long as word embeddings still identify words, the regression model will be able to relearn these (cross-space) stimulus dependencies. Correcting the embeddings through residualisation is, therefore, insufficient: As long as the signal being predicted contains the temporal structure inherent in naturalistic language, the proposed hallmarks will emerge.

Finally, it is important to note that in order to ensure that these findings are not unique to one specific stimulus set, we applied the same control analyses to the stimulus material from the original study by Goldstein et al. (2022b), and consistently observed the same effects: Proposed hallmarks of prediction emerged in both control systems (Figures S2B, S3B) and could not be corrected for by either bigram removal (Figure S5C) or residualisation (Figure S6B). Our results, therefore, demonstrate that the proposed hallmarks of prediction observed in neural encoding studies can be explained fully by stimulus dependencies, not only in the MEG datasets analysed here, but in the very dataset used to establish pre-onset encoding as evidence for neural prediction.

Discussion

We evaluated an encoding modeling paradigm that uses vector representations of words to capture neural signatures of linguistic predictions during naturalistic listening. Specifically, we asked to what extent two proposed hallmarks of prediction—namely (i) positive encoding performance prior to word onset, and (ii) sensitivity to the predictability of a word—can be interpreted as reflecting neural pre-activation. Across two MEG datasets, we found that both hallmarks appear not only in neural data, but also in two passive control systems: static word embeddings and the stimulus acoustics. Since these control systems encode the stimulus but cannot predict upcoming words, these results demonstrate that the proposed hallmarks can arise from stimulus dependencies alone, without assuming any predictive pre-activation in the brain. We further showed that methods proposed to correct for these dependencies—removing reoccurring bigrams or residualising neighbouring word information from embeddings—are insufficient, since pre-onset encoding persists in the acoustic control system even after such corrections. These results reveal a fundamental ambiguity in the approach. Both proposed hallmarks can be fully accounted for by stimulus dependencies alone, without assuming any prediction in the brain.

We observed that the first hallmark of prediction—pre-onset encoding—is tremendously robust, replicating not only in non-invasive MEG data (which has a lower signal-to-noise ratio compared to the ECoG data used by Goldstein et al. (2022b)), but also across various types of word embeddings (GPT-2, GloVe and even arbitrary embeddings), data sets (single-subject as well as multi-subject), MEG spaces (source as well as sensor data) and type of linguistic representation (neural, artificial or acoustic). By contrast, the second hallmark—i.e., the modulation of pre-onset encoding by next-word predictability—could only be replicated reliably in MEG in the few-subject dataset (Figure 1C, 3C, S1A-B, S7A-C), which is the same dataset analysed by Azizpour et al. (2024). We see two potential explanations for this discrepancy. First, this may simply be a result of differences in the amount of data, given that the single-subject dataset comprised more than ten times the amount of data per subject than was available in the multi-subject dataset. Alternatively, and more convincingly, this discrepancy could result from a difference in the experimental designs: the multi-subject dataset did not use natural speech but carefully

manipulated computer-generated speech, to minimise acoustic confounds such as co-articulation (see Methods for more details). Indeed, the second hallmark was not only absent in the neural data, but critically, also in the acoustics of the multi-subject dataset (see Figure S3A and S6A). Hence, in both datasets, neural encoding closely mirrored acoustic encoding results, suggesting that ostensible hallmarks of prediction observed in the neural data reflect the correlation structure of the stimulus material rather than neural pre-activation in the brain. This apparent discrepancy, therefore, supports our proposition that stimulus dependencies constitute the driving factor behind brain encoding results.

Another important finding of the present paper is the difficulty of removing or correcting for such correlations in the stimulus material. Natural language is rife with temporal structure that is useful for predicting neighbouring words—whether such structure may be semantic, syntactic, acoustic, or due to n-gram statistics. While residualisation successfully removes dependencies within a single representational space (Figure 4B), brain encoding involves a mapping *between* spaces: from word embeddings to neural responses. Removing temporal correlations from but one of these two spaces, still allows for the regression model to capitalise on the regularities and correlations in the second representational space. This is exemplified by our acoustic encoding analysis: even residualised word embeddings can predict the acoustics of preceding words, since the dependencies exist in language itself, not just in any particular representation of it.

Critically, we do not want to suggest that our results question the role of prediction during language processing itself. Indeed, there is a large body of work suggesting that human language processing is inherently predictive. For instance, readers and listeners are highly sensitive to even subtle fluctuations in linguistic predictability (Smith and Levy, 2013; Frank et al., 2015; Willems et al., 2016; Shain et al., 2024; Heilbron et al., 2023; Brodbeck et al., 2022; Heilbron et al., 2022; Szewczyk and Federmeier, 2022). However, such surprisal-based predictability effects on brain responses to language are usually post-stimulus (and hence indirect), while pre-stimulus evidence has, for some researchers, been considered the 'gold standard' in evidence for linguistic prediction (Kuperberg and Jaeger, 2016; Pickering and Gambi, 2018; Nieuwland, 2019). On first consideration, encoding modeling provides a new and more direct line of evidence that can assess pre-stimulus prediction (Schrimpf et al., 2021; Goldstein et al., 2022b; Caucheteux et al., 2023; Azizpour et al., 2024). However, due to the opacity of the word embeddings and regression models involved, interpreting these results as evidence for prediction in the brain is challenging (Antonello and Huth, 2024; Azizpour et al., 2024), rendering the evidence less direct than it may initially appear, even when it concerns evidence of pre-stimulus brain activity.

While these results, taken together, pose challenges to using pre-stimulus brain encoding to test for neural pre-activation, we yet see two possible future applications to use this analytical framework for this purpose. First, the predictability of the stimulus could be used as a threshold or benchmark: If brain encoding prior to word onset is quantitatively higher than the predictability of the stimulus, this might indicate that a predictive representation adds to the encoding performance stemming from temporal correlations alone. Although this is not the case in this current study, MEG is limited in terms of signal-to-noise ratio, and less noisy data, such as ECoG, might be able to fulfill this criterion. Secondly, stimulus predictability could be used as a tool in order to pre-select trials in which stimulus correlations do not allow for pre-onset encoding or favour a different trend (as in the case of our multi-subject dataset).

While our analyses focused on linear regression-based *encoding* models, the same logic applies to decoding approaches. Goldstein et al. (2022b) also presented decoding evidence, using convolutional neural networks to decode word identity from pre-onset neural activity. However, the same fundamental ambiguity remains: If temporal dependencies enable a regression model to predict neural activity from future word embeddings, they equally enable a decoder to predict future word identity from current neural activity—as both exploit the same underlying correlations in the stimulus material. Indeed, more powerful non-linear decoders may only amplify the problem as they learn to exploit subtler dependencies that linear models might miss. The control system logic applies here as well: If a decoder can predict upcoming words from pre-

onset acoustics, this would demonstrate that decoding performance need not reflect neural pre-activation. In other words, if the brain generates predictive pre-activations, this should increase mutual information between pre-onset neural activity and the upcoming word beyond what is present in the stimulus itself

Embedding-based encoding and decoding analyses represent an enticing new approach to studying language processing. However, their reliance on statistical prediction of brain activity, paradoxically, renders these methods difficult to use in order to test for predictions in brain activity. Since the same regularities that the brain may use to predict natural language can also be exploited by the regression model, it is ultimately difficult to know which system is performing the prediction—the brain or the regression model used by the researchers.

Methods

Data

To test whether previously observed evidence for encoding a word's pre-activation in neural data can be can be replicated in non-invasive MEG data, we used a publicly available, high quality MEG data set (Armeni et al., 2022) in which three subjects listened to the audiobook version of the entire Sherlock Holmes corpus in ten one-hour-long recording sessions. To minimise noise, head motion was restricted using individual 3D-printed head casts (for details on the data set and stimuli see Armeni et al., 2022). Brain responses were source localised and minimally filtered between 0.1 – 40Hz. For the analysis, the MEG data was time-locked to word onset from −2000*ms* to +2000*ms* without applying any baseline correction. Subsequently—akin to the analysis performed by Goldstein et al.—neural activity was averaged over a sliding window of 100*ms* which moved with a step size of 25*ms*. This resulted in 85 719 trials of 157 time points within the window of 4*s*.

To examine whether our findings replicate in a more conventional multi-subject data set, we repeated our analyses in another publicly available MEG data set with 27 participants (see Gwilliams et al., 2022 for more details). Crucially, the data set differed in several important aspects from the main data set used in this analysis. Firstly, participants listened to 4 different stories within their one hour-long recording session, resulting in a total of 7 745 trials per participant. Secondly, listening was interrupted by a word list or question every 3 minutes. Thirdly, the speech rate varied every 5 – 20 sentences between 145 and 205 words per minute and silences between sentences varied from 0 – 1000*ms*. And lastly, MEG data was not source-localised, but all analyses were performed on channel data.

MEG encoding modeling

The neural response to any given word was predicted based on word embeddings by means of a ridge regression—separately for each source and time point—within a 10-fold cross-validation. Within each fold, the features of the models were standard scaled before running the regression. For each time point, the predicted response was then correlated with the actual response (see Figure 1 A for a visualisation). As features in the regression model, embeddings were obtained for each word.

For the non-contextualised analysis, we used 300-dimensional GloVe vectors (Pennington et al., 2014) obtained from the *spacy* package, version 3.4.3 (Honnibal and Montani, 2017). For the contextualised analysis, 768-dimensional word embeddings were extracted from the 8^{th} layer of Huggingface's (version 4.23.1) pre-trained GPT-2-S (Radford et al., 2019), as middle layers have been shown to result in the best brain encoding performance (Goldstein et al., 2022a; Caucheteux and King, 2020; Schrimpf et al., 2020). For words which consisted of multiple byte-pair encoded tokens (BPEs), such as "Sherlock" which is broken down into "S-her-lock", the embedding of the last BPE was used. For computational reasons, the contextual window for each word ranged from 512 to 1024. For the arbitrary model, 300-dimensional word-specific vectors were drawn from a Gaussian distribution (M = 0.1, SD = 1.1). Hence, arbitrary vectors contained no systematic information other than word identity.

In order to test the second hallmark, i.e., the sensitivity of the encoding performance to the predictability of a word, we split the data into easily predictable and less predictable words, ran a separate encoding model for each split and compared their encoding performance. Since GPT-2's internal word representations are a combination of previous word representations, encoding results are difficult to interpret temporally. Consequently, this analysis was performed for non-contextualised GloVe embeddings only. In order to ascertain whether the encoding performance for predictable words was significantly larger than for unpredictable words, we performed cluster-based permutation testing using threshold-free cluster enhancement (TFCE) with 10 000 permutations as implemented in mne stats' permutation_cluster_1samp_test function. Differences were considered significant if computed t-value exceeded the 95-th percentile under the permutation distribution.

Words were defined as easily predictable as opposed to less predictable words based on GPT-2-XL's top-one prediction. This resulted in 30 590 predicted and 55 129 unpredicted words for the few-subject dataset and in 2 124 predicted and 5 621 unpredicted words for the few-subject dataset. Given that GPT-2-XL's top-one prediction might constitute a conservative estimate of whether or not a word was predictable in context, we repeated this analysis but defined words as easily predictable if they were among GPT-2-XL's top-five predicted words. This resulted in 50 666 predicted and 35 053 unpredicted words for the few-subject dataset and in 3 703 predicted and 4 042 unpredicted words for the multi-subject dataset. Results for these different splits are shown in the supplement.

Source Selection

Given that the purpose of this study was to encode neural responses related to linguistic predictions, we aimed to restrain our model to sources related to language processing. Hence, sources were selected for each subject individually according to a 2-step procedure. First, prior to the encoding modelling, we pre-selected sources based on whether they were located in the bilateral language network (see Heilbron et al., 2022). This resulted in 100 sources, which were used for the encoding model. The data matrix, therefore, had the shape 100 × 85 719 × 157. Out of the resulting 100 encoded sources, we retained only those per subject which proved to allow for good encoding performance post word onset.

For this purpose, we determined for each subject the peak encoding performance in the postword-onset window of 0 − 500*ms*. We subsequently defined a cut-off threshold of what constitutes a "good encoding performance" at least 30% of that peak value. A source was then considered to allow for good encoding if it reached this threshold within the post-word-onset window of 0−500*ms*. We chose this time-window for our selection process, since encoding related to the word itself (as opposed to other spurious elements in the data) should be highest while the word is perceived and processed, i.e. during a time-window when well known components such as the N400 or P600 are usually observed. Additionally, both hallmarks of prediction concern the encoding performance prior to word onset since they are supposed to reflect an encoding of the pre-activation of the representation of a word. Hence, selecting sources based on post-onset encoding performance avoids double dipping.

Source selection was based on GloVe encoding results. This procedure resulted in 32 sources for subject 1 (max = 0.150, threshold = 0.045), 33 sources for subject 2 (max = 0.103, threshold = 0.031), 25 sources for subject 3 (max = 0.187, threshold = 0.056). To ensure that source selection was stable across neural network representations, we performed the same procedure based on GPT-2 encoding results. This resulted in fewer sources (26, 22, and 20 sources for subject 1, 2 and 3 respectively), all of which were a subset of the sources obtained from the GloVe based selection. Hence, all analyses were performed with the GloVe based selection in order to ensure greater inclusivity.

Since analyses in the multi-subject data were performed on channel and not source-localised data, and since encoding was performed on group-level, channel selection was based on a simple common cut-off threshold. Hence, the encoding model was run on all 208 channels for each subject separately, resulting in 27 data matrices of the size of 208 × 7 745 × 157. For each

participant in the dataset, channels were retained for plotting if the channel reached the threshold in the post-word-onset window of 0 – 500*ms*. This threshold was selected based on our results from the few-subject analysis (threshold = 0.031). This resulted in the exclusion of subject 12 for whom no channel surpassed the threshold.

Control system one: self-predictability analysis

As mentioned above, due to the inherent structure present in natural language, neighbouring words can share information, and therefore, neighbouring word representations can be correlated. For instance, nouns are frequently preceded by articles or prepositions, and neighbouring words belong to a similar semantic field ("pine tree", "driving a car", etc.) Hence, a positive encoding performance prior to word onset might result from neighbouring embeddings being correlated, and each embedding encoding the neural representation of the corresponding word, not due to a pre-activation in the neural signal. In other words, if a word's representation $\boldsymbol{x}_0$ and the representation of the preceding word $\boldsymbol{x}_{-1}$ are correlated ($\rho(\boldsymbol{x}_{-1}, \boldsymbol{x}_0) > 0$) and each representation ($\boldsymbol{x}_{-1}, \boldsymbol{x}_0$) successfully encodes it's corresponding neural activity ($\boldsymbol{y}_{-1}, \boldsymbol{y}_0$), then encoding prior to word onset is possible since $\boldsymbol{x}_0$ can be used to approximate $\boldsymbol{x}_{-1}$, leading to a lower, but positive encoding performance. Word vector representation, therefore, constitute a meaningful control system, since stimulus dependencies present in natural language are encoded within them, while they do not perform any active prediction themselves.

In order to investigate this possibility, we constructed an encoding model in which the dependent variable, i.e. the y-matrix for each trial, did not consist of neural data, but of the embedding vectors of the words that were presented at each time point. For instance, given the sentence "You know my methods, Watson." time-locked to the onset of the word "methods", we computed the onset and offset times of each of our 157 sliding time points, determined which word was presented at that time point and filled that data point with the vector for that word. For example, if we assume that each word above had a duration of 500*ms*, the 20 time points between −1500 and −1000*ms* were filled with the vector for "You", the next 20 time points with the vector for "know" etcetera (see Figure 2A). Due to computational reasons self-predictability was only computed for the first session in the few-subject dataset, i.e. approximately 10% of the available data (8 622 words). We deemed this sufficient since the stimulus material in the few-subject dataset consisted of one text corpus, namely the full Sherlock Holmes corpus, and therefore the correlational structure in 10% of the data might reasonably be representative for the whole text. Additionally, unlike the MEG data in our brain encoding model, Thus for GPT-2, the dependent variable was a 768 × 8 622 × 157-dimensional matrix, and for GloVe and arbitrary vectors it was 300 × 8 622 × 157-dimensional.

Akin to the brain encoding, modeling was performed by means of a 10-fold cross-validated ridge regression in order to predict previous embeddings from the embedding at time point zero, and correlated the predicted and actual embeddings. This regression was performed for each feature and time point separately and both the y- and the X-matrix were standard scaled within each fold. The resulting correlation will be referred to as the *self-predictability* of a model.

In order to test whether model self-predictability would also be able to account for the second proposed hallmark of prediction, namely sensitivity to next word predictability, we repeated the same procedure as for the neural data. To compare the self-predictability of GloVe vectors of predictable as opposed to less predictable words, we split the data again based on GPT-2-XL's top-one prediction (see supplement for results from splits based on GPT-2-XL's top-five prediction). This split resulted in 3 075 correctly and 5 547 incorrectly predicted trials (5 054 and 3 568 for correct and incorrect predictions in the top-five split), thereby closely mirroring the percentages from the whole data set. Since our multi-subject dataset consisted of merely 7 445 trials, the entirety of the data was used for the self-predictability analysis. Hence, the splits resulted in 2 124 predicted and 5 621 unpredicted words (for the top-one split) and in 3 703 predicted and 4 042 unpredicted words (for the top-five split).

In order to ensure that the results from our control analysis are not specific to the stimulus material of our two MEG datasets, but generalise to the original study that first proposed the two hallmarks of prediction tested here, we applied our control analyses also to the stimulus material used by Goldstein et al. (2022b). In their study, the authors presented their participants with the first 30 minutes of the episode *Monkey in the Middle* of the *This American Life* podcast, resulting in a total of 5 136 words. Since the authors used the last layer, i.e., layer 47, of GPT-2-XL in their original study, we also used their published GPT embeddings for the analysis instead of using layer 8 of GPT-2-S. Hence, the dependent variable, the y-matrix, was a 1 600 × 5 136 × 157-dimensional matrix, instead of being 768-dimensional as in the case of our few- and multi-subject datasets. The predictability split resulted in 1 485 predicted and 3 651 unpredicted words (for the top-one split) and in 2 604 predicted and 2 532 unpredicted words (for the top-five split).

Control system two: acoustic encoding analysis

Our first control system used word embeddings (GPT, GloVe and arbitrary representations) in order to predict word embeddings that had occurred up to 2*s* prior to that word. Hence, the mapping performed by the regression model occurred within one representational system. Brain encoding designs, however, involve a mapping from one feature space, e.g. word embeddings, to an entirely different representational system, i.e. MEG or ECoG responses. In order to ascertain to what extent temporal dependencies in the stimulus material, might still be able to explain ostensible hallmarks of prediction, even when mapping between representational spaces, we analysed the predictability of the acoustics of each word. We deem word acoustics to constitute a meaningful, passive control system, as they must be represented in the brain of our listeners, and yet they in themselves quite obviously do not perform any active prediction. We obtained an acoustic embedding of each word by computing the average 8-Mel spectrogram and envelop of that word. Like for the self-predictability analysis, we constructed acoustic epochs by computing the onset and offset times of each of our 157 time points within the 4*s*interval, determined which word was presented at that time point, and filled that data point with the acoustic representation of that word. This resulted in a y-matrix of 85 719 × 9 for the few-subject dataset, 7 745 × 9 for the multi-subject dataset, and 5 136 × 9 Goldstein et al. (2022b)'s podcast dataset. We then tested the predictability of this representation of our stimulus material, and it's modulation by next-word predictability using the same encoding model as used for the brain encoding and self-predictability analysis (see Figure 3A).

Note that for the multi-subject dataset no information about the offset of a word was available, hence we used the onset of the next word as a proxy for a word's offset. This however, meant that in our multi-subject dataset the acoustic representation for each word contained silences occurring between words and sentences. This—with no doubt—resulted in poorer acoustic embeddings and might have driven the differences in time course, and the lack of a predictability effect observed in the acoustic encoding of this dataset.

Crucially, in Goldstein et al. (2022b)'s podcast data, the words preceding predictable or unpredictable words, i.e. the words that had to be predicted by the regression model, differed substantially in their part-of-speech (PoS) tags, with more content words for predictable trials (52% vs. 45%) and more function words for unpredictable trials (55% vs. 48%). Crucially, this was not the case for our few subject dataset in which content and non-content words were almost perfectly balanced between splits, at 46% and 54% respectively. Hence, since content and non-content words differ in length and acoustic variability, and this acoustic variability of highly frequent function words has been found to depend on their predictability in context Bell et al. (2003), we expected these differences to be a potential confounding factor in the encoding performance. Furthermore, we expected differences in the amount of trials between the splits to affect encoding performance, especially in low data regimes. To account for these potential confounds in the acoustics of Goldstein et al. (2022b)'s podcast data, we randomly sub-sampled the predictable split to resemble the PoS distribution of the overall text without eliminating bigrams. We then sub-sampled the unpredictable split to resemble the resulting PoS distribution of the predictable split, while keeping the number of trials in each split of similar magnitude. This

resulted in 857 predictable and 792 unpredictable trials for the top-one split, and 1 554 predictable and 1 480 unpredictable trials for the top-five split. We repeated this procedure for 100 random seeds to avoid artifacts caused by a specific random draw.

Accounting for stimulus dependencies

In order to assess whether methods that have been proposed to correct for stimulus dependencies truly eliminate their influence on the hallmarks of prediction put forth by Goldstein et al. (2022b), we applied two possible approaches—removing reoccurring bigrams and vector residualisation— to both our control systems.

Removal of reoccurring bigrams

The first approach for controlling for stimulus dependencies we tested was the removal of all reoccurrences of bigrams from our data. Hence, for any bigram of words occurring in the stimulus material we only retained its first occurrence. This resulted in a reduction of data to 43 670 (from 85 719) for the few-subject dataset, to 6 566 (from 7 445) for the multi-subject dataset, and to 4110 (from 5 136) for the podcast dataset. We re-tested for positive pre-onset encoding performance—especially focusing on our arbitrary model—within each of our two control systems. If removing all reoccurring bigrams can account for temporal dependencies in the stimulus set, using arbitrary vectors in which nothing but word identity is encoded, should not result in any positive pre-onset encoding in our passive control systems.

Residualisation of word embeddings

As mentioned above, correlations between neighbouring, structured word embeddings (such as GloVe and GPT-2 representations) might explain positive pre-onset encoding results. Since neighbouring word embeddings are positively correlated, the word embedding of the word in question (w_0) can be used to predict the preceding word (w_{-1}), and can therefore, be used in order to predict brain activity prior to word onset without any predictive pre-activation in the neural signal. In this second approach, we therefore, aimed to determine whether correcting for this correlation within the word embeddings is sufficient to account for the temporal dependencies in the stimulus material. For this purpose, word embeddings were residualised. For each word in the text, we regressed-out all the information of the embedding at time point zero ($\boldsymbol{x}_0$) that could be linearly predicted from the preceding embedding ($\boldsymbol{x}_{-1}$). Hence, given the sentence “You know my methods.”, “You” is regressed out of “know”, “know” is regressed out of “my”, and “my” out of “methods”. This analysis can be seen as a generalised version of the control analysis performed by Goldstein et al. (2022b), based on directly projecting out neighbouring word embeddings. We then re-tested for positive pre-onset encoding and its modulation by predictability using residualised word embeddings in both our control systems as well as the two MEG datasets.

Supplementary Figures

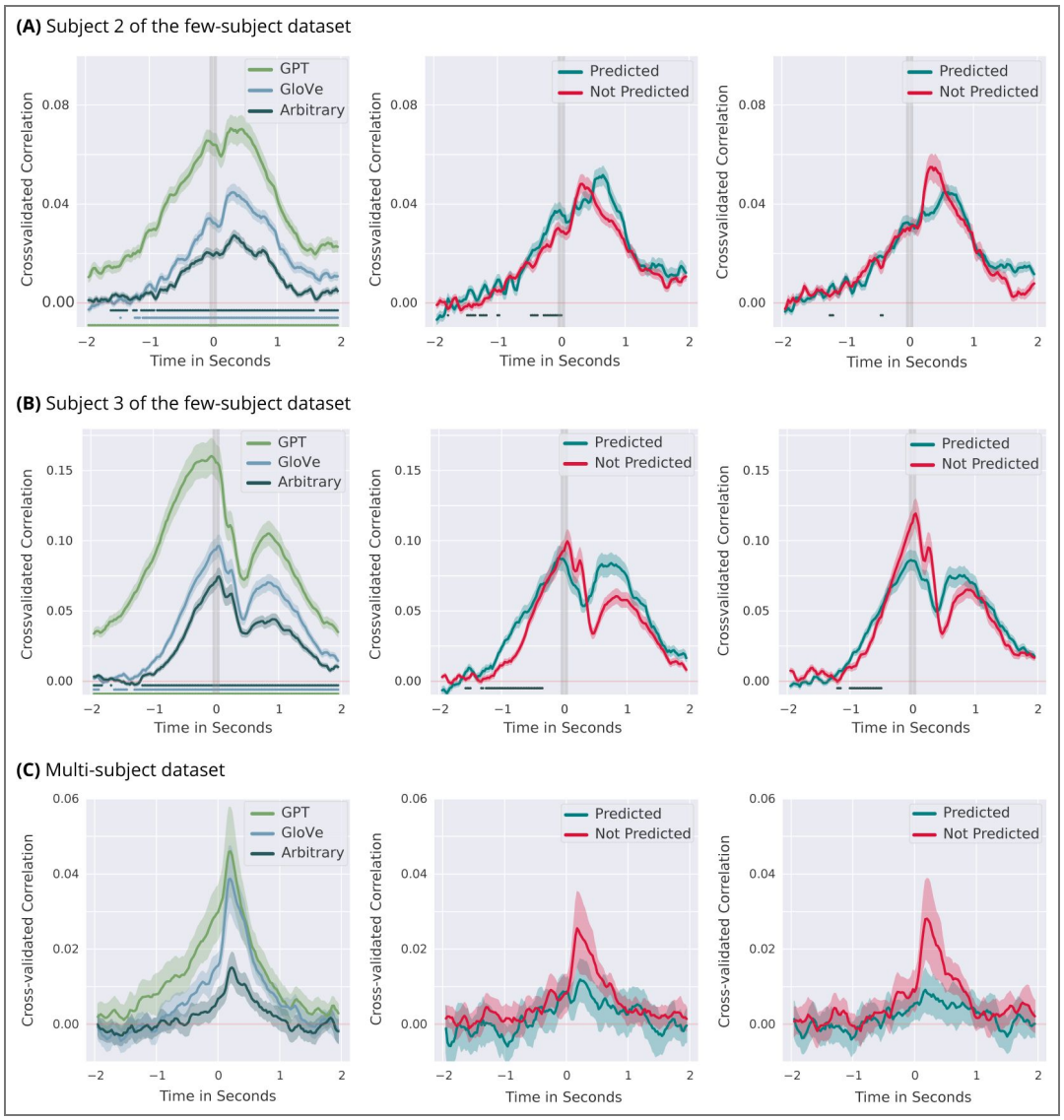


Figure S1. Hallmarks of prediction in the remaining two subjects of the few-subject dataset and the Multi-subject dataset. First panel from the left shows the overall encoding performance of GPT-2, GloVe and arbitrary vectors. Lines show clusters of time points for which encoding performance is significantly different from zero ($p < 0.05$ under the permutation distribution). Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-one prediction and the right panel shows the the sensitivity to GPT-2-XL's top-five prediction. Lines show clusters of time points for which encoding performance is significantly larger for the predictable as opposed to unpredictable words prior to word onset ($p < 0.05$ under the permutation distribution). Shaded areas show 95-% confidence intervals computed over sources and cross-validation splits in the single-subject analyses and over subjects in the multi-subject analysis. For the multi-subject dataset we find no evidence for the second hallmark of prediction, i.e. no sensitivity to the predictability of a word for pre-onset encoding performance.

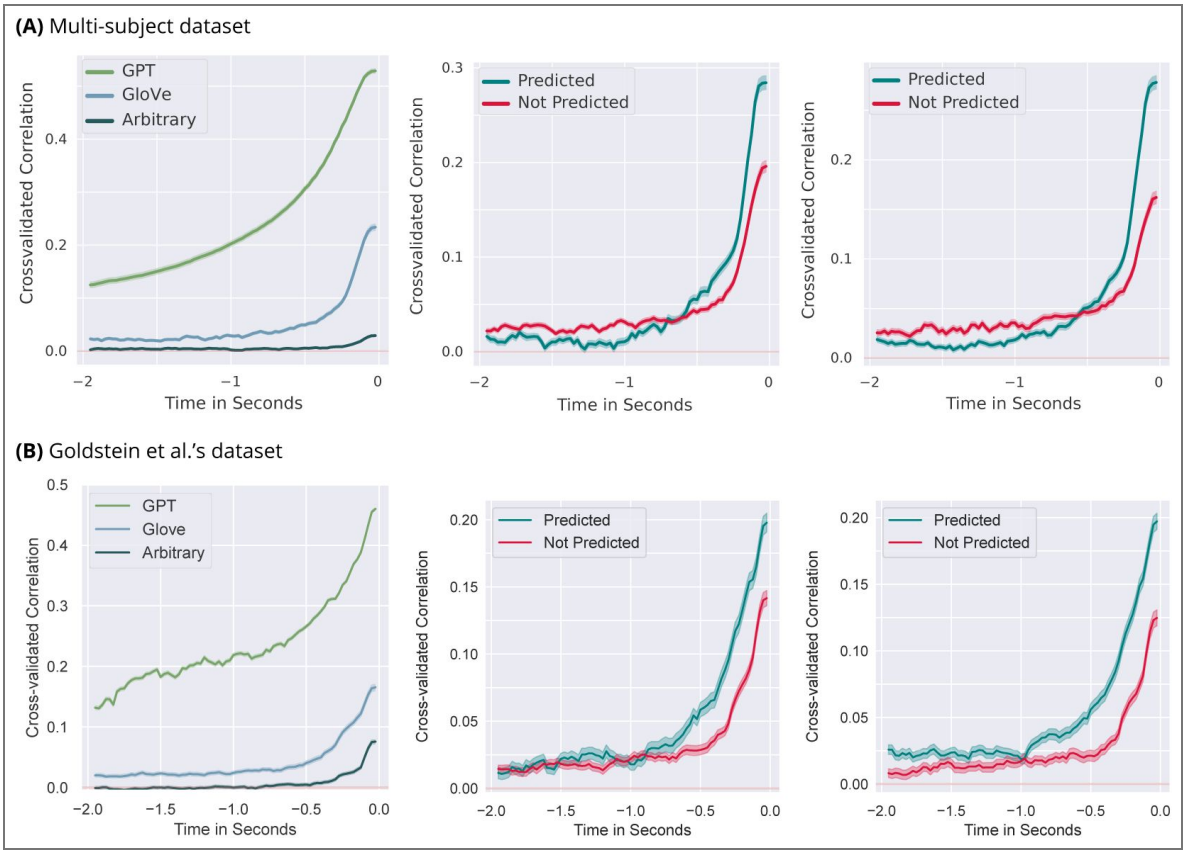


Figure S2. Self-Predictability in the Multi-subject dataset and in Goldstein et al. (2022b) data.

First panel from the left shows self-predictability of GPT-2, GloVe and arbitrary models. Middle panel shows the sensitivity of self-predictability of GloVe to the predictability of the word as defined by GPT-2-XL's top-one prediction and the right panel shows the the sensitivity as defined by GPT-2-XL's top-five prediction. Shaded areas show 95-% confidence intervals computed over model dimensions. Self-predictability results are identical to those observed in the few-subject dataset. We find both ostensible hallmarks of prediction in the stimulus material, namely the word embeddings of the material.

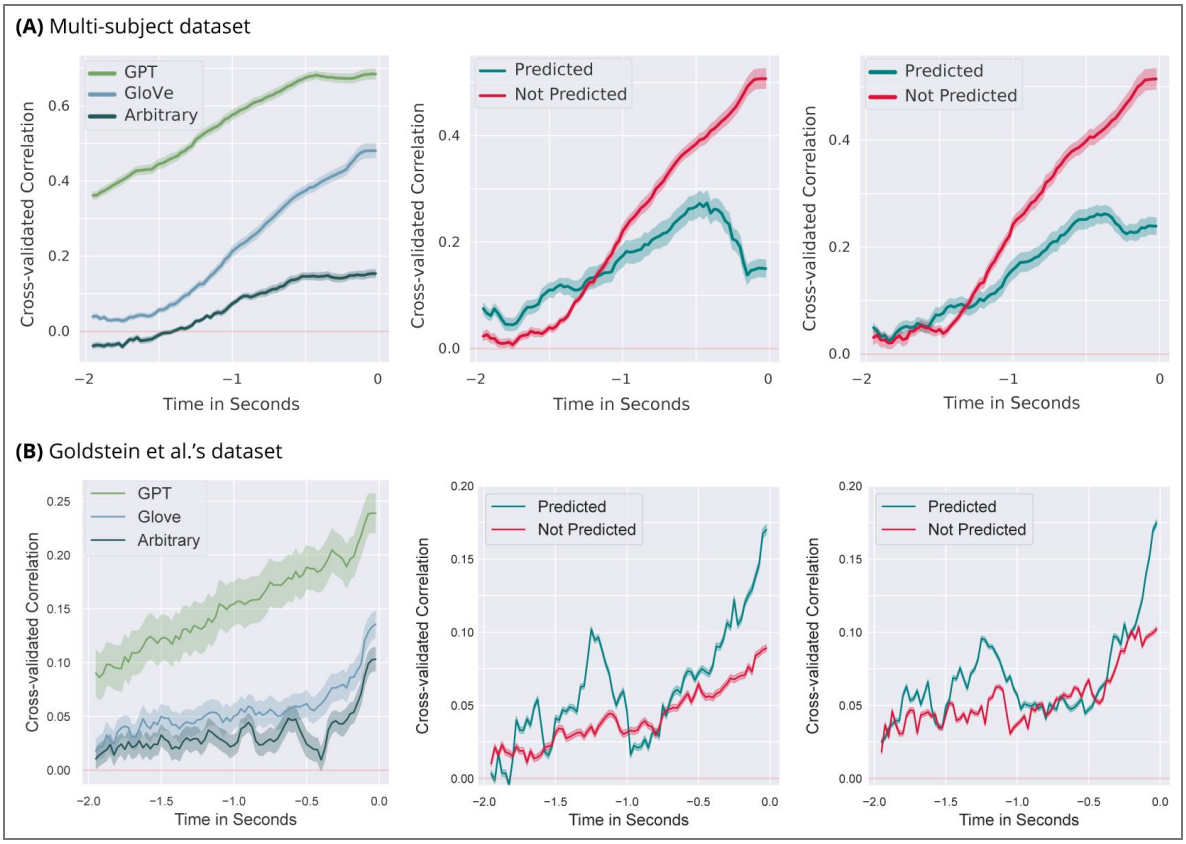


Figure S3. Predicting Acoustics prior to word onset from *original* embedding vectors for both datasets.

First panel from the left shows the overall encoding performance of residualised GPT-2, GloVe and arbitrary vectors. Results closely mirror brain encoding results both in encoding time course and differences between models (see Figure 1B and C and S1). Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-one prediction and the right panel shows the the sensitivity to GPT-2-XL's top-five prediction when using original GloVe embeddings. Shaded areas show 95-% confidence intervals computed over the 9 dimensions (8 mels + envelope). Results mirror qualitative differences observed in the brain encoding studies.

Figure S4. Encoding data using GPT, GloVe and arbitrary vectors after removing re-occurrences of bigrams, i.e. only retaining the first occurrence, in our few-subject dataset.

Panel A) shows reduced but significant pre-onset encoding for all three vectors of the MEG data (subject 1). Panel B) shows model self-predictability after removing re-occurring bigrams. Indeed, removing bigrams had no effect on the self-predictability of any of the three models. Panel C) shows the pre-onset predictability of the acoustics of our few-subject dataset after removing re-occurring bigrams. As for the self-predictability, removing bigrams led to almost identical encoding performance as in S6A). This shows that removing reoccurring bigrams does not account for dependencies in the stimulus material and pre-onset encoding of the neural data might be driven by the predictability of stimulus or model features prior to word onset. Shaded areas show 95-% confidence intervals computed over 10 cross-validation folds and the dimensionality (for the selfpredictability and acoustics) and channels (for the MEG data). Results mirror qualitative differences observed in the brain encoding studies.

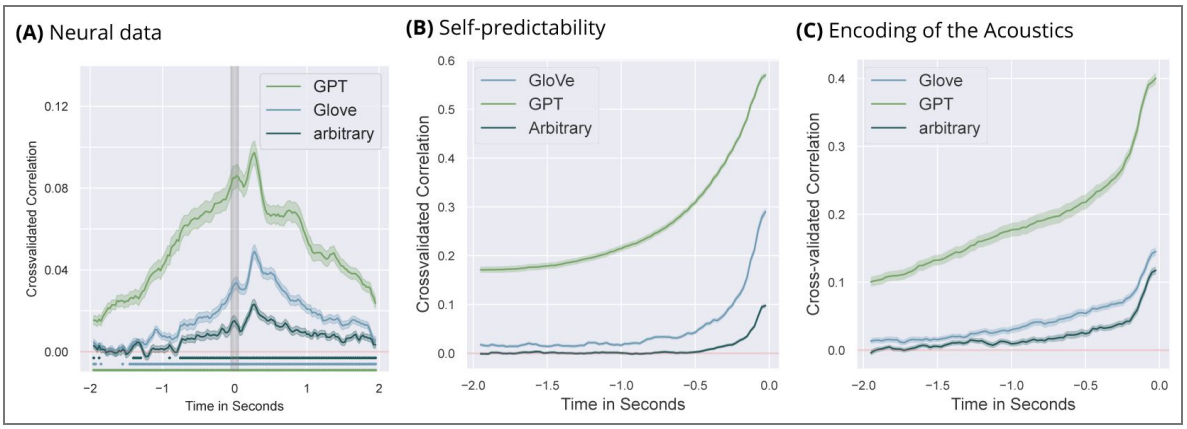
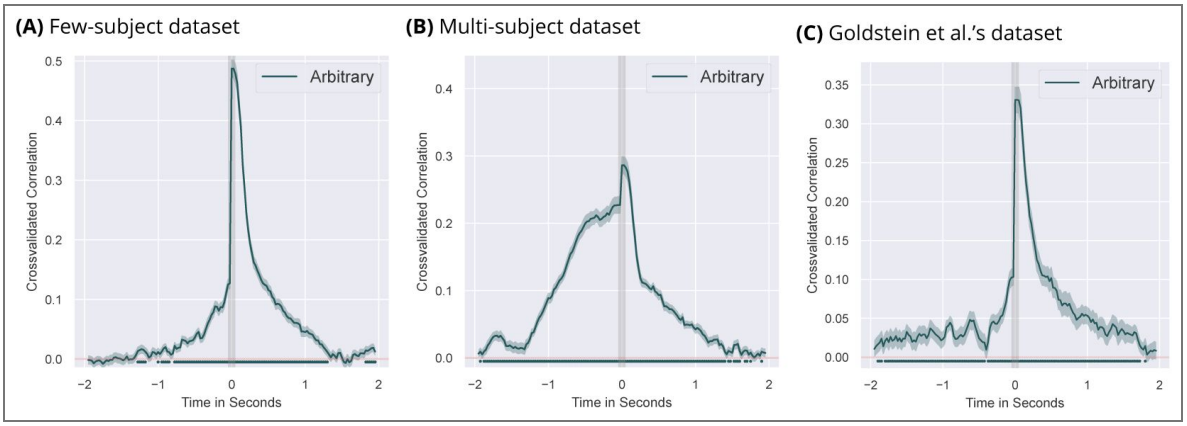


Figure S5. Predicting Acoustics from arbitrary embedding vectors for all three datasets after all reoccurring bigrams have been removed.

Results show that even when reoccurrances of bigrams are removed, it is still possible to find significant pre-onset encoding for arbitrary vectors which is solely due to encoding temporal dependencies in the stimulus material. Shaded areas show 95-% confidence intervals computed over 10 cross-validation folds and the 9 dimensions of our acoustic data (8 mels + envelope). Results mirror qualitative differences observed in the brain encoding studies.



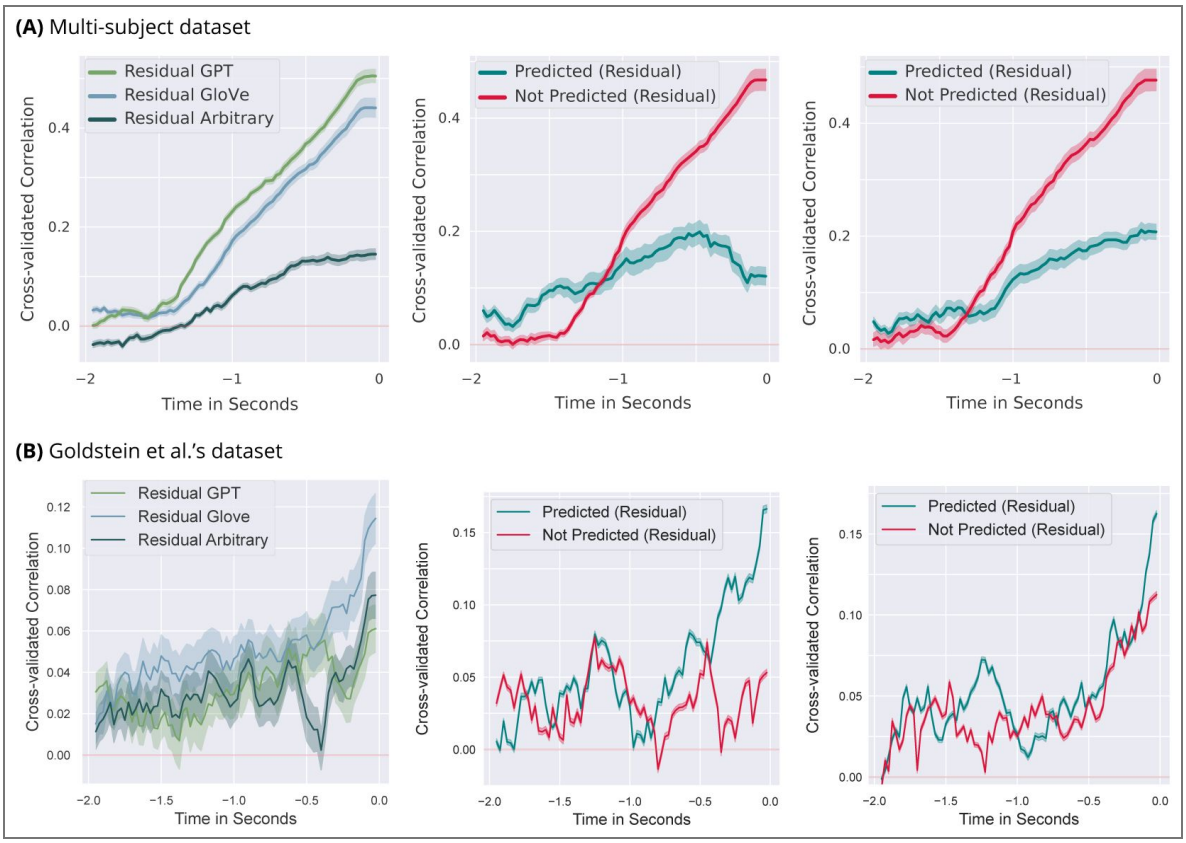


Figure S6. Predicting Acoustics prior to word onset from *residual* word embeddings in the multi-subject dataset.

First panel from the left shows the overall encoding performance of residualised GPT-2, GloVe and arbitrary vectors. Results mirror brain encoding results both in encoding time courses and differences between models (see Figure S7 C). Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-one prediction and the right panel shows the the sensitivity to GPT-2-XL's top-five prediction when using residualised GloVe embeddings. Shaded areas show 95-% confidence intervals computed over the 9 dimensions (8 mels + envelope). Results mirror qualitative differences observed in the brain encoding study.

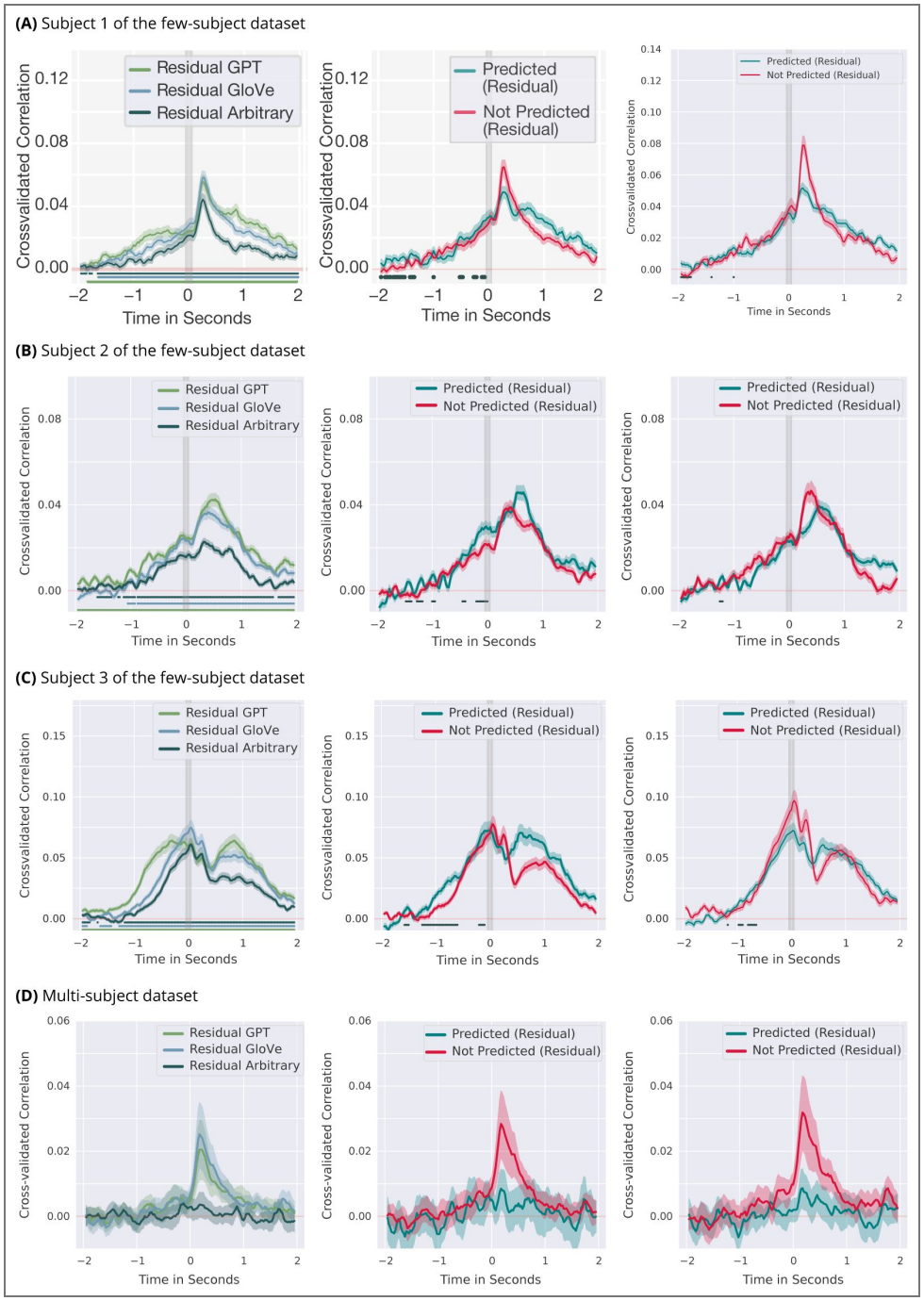


Figure S7. Ostensible hallmarks of prediction after removing model self-predictability through residualising word embeddings in the few-subject dataset and the multi-subject dataset.

First panel from the left shows the overall encoding performance of residualised GPT-2, GloVe and arbitrary vectors. Lines show clusters of time points for which encoding performance is significantly different from zero ($p < 0.05$ under the permutation distribution). Encoding performance as well as differences between models are reduced compared to encoding results with original vectors. Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-one prediction and the right panel shows the the sensitivity to GPT-2-XL's top-five prediction. Lines show clusters of time points for which encoding performance is significantly larger for the predictable as opposed to unpredictable words prior to word onset ($p < 0.05$ under the permutation distribution). Shaded areas show 95-% confidence intervals computed over sources and cross-validation splits in the single-subject analyses and over subjects in the multi-subject analysis. For the multi-subject dataset we find no evidence for the second hallmark of prediction, i.e. no sensitivity to the predictability of a word for pre-onset encoding performance.

Data availability

The code used for modelling analyses and plotting is available at *https://github.com/InesSchoenmann/Lingpred*.

Additional information

Funding

Funder	Grant reference number	Author
Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)	VI.C.231.043	Floris P de Lange
EC | European Research Council (ERC)	https://doi.org/10.3030/101000942	Floris P de Lange Micha Heilbron

Author ORCID iDs

Floris P de Lange: https://orcid.org/0000-0002-6730-1452
Micha Heilbron: https://orcid.org/0000-0003-3039-4007

References

Antonello R, Huth A (2024) Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language* **5**:64-79

Armeni K, Güçlü U, van Gerven M, Schoffelen JM (2022) A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data* **9**:278

Azizpour S, Westner BU, Szewczyk J, Güçlü U, Geerligs L (2024) Signatures of prediction during natural listening in MEG data?. *arXiv*

Bell A, Jurafsky D, Fosler-Lussier E, Girand C, Gregory M, Gildea D (2003) Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the acoustical society of America* **113**:1001-1024

Boston MF, Hale JT, Vasishth S, Kliegl R (2011) Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes* **26**:301-349

Brodbeck C, Bhattasali S, Heredia AAC, Resnik P, Simon JZ, Lau E (2022) Parallel processing in speech perception with local and global representations of linguistic context. *eLife* **11**:e72056 https://doi.org/10.7554/eLife.72056

Caucheteux C, Gramfort A, King JR (2023) Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour* 1-12

Caucheteux C, King JR (2020) Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*

Caucheteux C, King JR (2022) Brains and algorithms partially converge in natural language processing. *Communications biology* **5**:134

Frank SL, Otten LJ, Galli G, Vigliocco G (2015) The ERP response to the amount of information conveyed by words in sentences. *Brain and language* **140**:1-11

Goldstein A, Ham E, Nastase SA, Zada Z, Grinstein-Dabus A, Aubrey B, Schain M, Gazula H, Feder A, Doyle W, *et al.* (2022a) Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *bioRxiv*

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, *et al.* (2022b) Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* **25**:369-380

Gwilliams L, Flick G, Marantz A, Pylkkanen L, Poeppel D, King JR (2022) MEG-MASC: a high-quality magnetoencephalography dataset for evaluating natural speech processing. *arXiv*

Hale J (2001) A probabilistic Earley parser as a psycholinguistic model.

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, De Lange FP (2022) A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* **119**:e2201968119

Heilbron M, van Haren J, Hagoort P, de Lange FP (2023) Lexical processing strongly affects reading times but not skipping during natural reading. *Open Mind* **7**:757-783

Honnibal M, Montani I (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Sentometrics Research*

Huettig F (2015) Four central questions about prediction in language processing. *Brain research* **1626**:118-135

Huettig F, Mani N (2016) Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience* **31**:19-31

Jain S, Huth A (2018) Incorporating context into language encoding models for fMRI. *Advances in neural information processing systems* **31**

Kuperberg GR, Jaeger TF (2016) What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience* **31**:32-59

Kutas M, Hillyard SA (1980) Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology* **11**:99-116

Kutas M, Hillyard SA (1980) Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* **207**:203-205

Kutas M, Hillyard SA (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307**:161-163

Linzen T, Baroni M (2021) Syntactic structure from deep learning. *Annual Review of Linguistics* **7**:195-212

Manning CD (2022) Human language understanding & reasoning. *Daedalus* **151**:127-138

Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O (2020) Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* **117**:30046-30054

Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J (2024) Large language models: A survey. *arXiv*

Nieuwland MS (2019) Do 'early'brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews* **96**:367-400

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532-1543

Pickering MJ, Gambi C (2018) Predicting while comprehending language: A theory and review. *Psychological bulletin* **144**:1002

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. (2019) Language models are unsupervised multitask learners. *OpenAI blog* **1**:9

Schrimpf M, Blank I, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum J, Fedorenko E (2020) Artificial neural networks accurately predict language processing in the brain. *BioRxiv*

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E (2021) The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118**:e2105646118

Shain C, Meister C, Pimentel T, Cotterell R, Levy R (2024) Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences* **121**:e2307876121

Smith NJ, Levy R (2013) The effect of word predictability on reading time is logarithmic. *Cognition* **128**:302-319

Szewczyk JM, Federmeier KD (2022) Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of memory and language* **123**:104311

Tuckute G, Kanwisher N, Fedorenko E (2024) Language in brains, minds, and machines. *Annual Review of Neuroscience* **47**

Wang L, Kuperberg G (2018) Jensen O. Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *eLife* **7**:e39061 https://doi.org/10.7554/eLife.39061

Willems RM, Frank SL, Nijhof AD, Hagoort P, Van den Bosch A. (2016) Prediction during natural language comprehension. *Cerebral Cortex* **26**:2506-2516

Peer reviews

Reviewer #1 (Public review):

Summary:

This paper tackles an important question: What drives the predictability of pre-stimulus brain activity? The authors challenge the claim that "pre-onset" encoding effects in naturalistic language data have to reflect the brain predicting the upcoming word. They lay out an alternative explanation: because language has statistical structure and dependencies, the "pre-onset" effect might arise from these dependencies, instead of active prediction. The authors analyze two MEG datasets with naturalistic data.

Strengths:

The paper proposes a very interesting alternative hypothesis for claims in prior work (e.g., Goldstein et al., 2022). In contrast to claims in prior work, the current paper convincingly demonstrates that prior results can be explained by inherent stimulus dependencies in natural language, as opposed to the brain actively predicting future linguistic content.

Two independent datasets are analyzed. The analyses with the most and least predictive words is clever, and is nicely complementing the more naturalistic analyses. The work emphasizes how claims about linguistic prediction cannot be trivially drawn using encoding models in naturalistic designs.

https://doi.org/10.7554/eLife.106543.2.sa3

Reviewer #2 (Public review):

Summary:

At a high-level, the reviewers demonstrate that there is a explanation for pre-word-onset predictivity in neural responses that does not invoke a theory of predictive coding or processing. The paper does this by demonstrating that this predictivity can be explained solely as a property of the local mutual information statistics of natural language. That is, the reason that pre-word onset predictivity exist could simply boil down to the common prevalence of redundant bigram or skip-gram information in natural language.

Strengths:

The paper addresses a problem of significance and uses methods from modern NeuroAI encoding model literature to do so. The arguments, both around stimulus dependencies and the problems of residualization, are compellingly motivated and point out major holes in the reasoning behind several influential papers in the field, most notably Goldstein et al. This result, together with other papers that have pointed out other serious problems in this body of work, should provoke a reconsideration of papers from encoding model literature that have promoted predictive coding. The paper also brings to the forefront issues in extremely common methods like residualization that are good to raise for those who might be tempted to use or interpret these methods incorrectly.

Weaknesses:

After author revision, I see no major weaknesses in the underlying arguments or data processing steps.

https://doi.org/10.7554/eLife.106543.2.sa2

Reviewer #3 (Public review):

Summary:

The study by Schönmann et al. presents compelling analyses based on two MEG datasets, offering strong evidence that the pre-onset response observed in a highly influential study (Goldstein et al., 2022) can be attributed to stimulus dependencies-specifically, the auto-correlation in the stimuli-rather than to predictive processing in the brain. Given that both the pre-onset response and the encoding model are central to the landmark study, and that similar approaches have been adopted in several influential works, this manuscript is likely to be of high interest to the field. Overall, this study encourages more cautious interpretation of pre-onset responses in neural data, and the paper is well written and clearly structured.

Strengths:

• The authors provide clear and convincing evidence that inherent dependencies in word embeddings can lead to pre-activation of upcoming words, previously interpreted as neural predictive processing in many influential studies.

• They demonstrate that dependencies across representational domains (word embeddings and acoustic features) can explain the pre-onset response, and that these effects are not eliminated by regressing out neighboring word embeddings-an approach used in prior work.

• The study is based on two large MEG datasets and one ECoG dataset, showing that results previously observed in ECoG data can be replicated in MEG. Moreover, the stimulus dependencies appear to be consistent across the three datasets.

Weaknesses:

• While this study shows that stimulus dependency can account for pre-onset responses, it remains unclear whether this fully explains them, or whether predictive processing still plays a role. The more important question is whether pre-activation remains after accounting for these confounds.

Comments on revisions:

I appreciate the added analyses. This study raises an important methodological concern regarding an influential paper and will certainly have a high impact on our field.

https://doi.org/10.7554/eLife.106543.2.sa1

Author response:

The following is the authors' response to the original reviews

We thank the reviewers for their constructive feedback, which has helped preparing a substantially improved manuscript. In response to concerns about the conceptual distinction between prediction and stimulus dependency, we have fundamentally restructured the paper around the notion of passive control systems. This involved rewriting the Abstract, Introduction, and large portions of the Results (~60% of text revised).

Key changes:

- New analyses on Goldstein et al. (2022) data. We demonstrate that our findings—including the insufficiency of proposed corrections—generalise to the original dataset (Figures S2B, S3B, S5C, S6B).

- Clarified novel contribution. We now make explicit that prior control analyses (residualisation, bigram removal) do not address the concern, because hallmarks persist in passive systems that cannot predict.

- Proposed criterion for future work. Pre-onset neural encoding can only count as evidence for prediction if it exceeds a passive baseline (e.g., acoustics).

We believe the revision offers a clearer, more rigorous contribution and provides a constructive framework for evaluating claims of neural prediction.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *This paper tackles an important question: What drives the predictability of pre-stimulus brain activity? The authors challenge the claim that "pre-onset" encoding effects in naturalistic language data have to reflect the brain predicting the upcoming word. They lay out an alternative explanation: because language has statistical structure and dependencies, the "pre-onset" effect might arise from these dependencies, instead of active prediction. The authors analyze two MEG datasets with naturalistic data.*
>
> *Strengths:*
>
> *The paper proposes a very reasonable alternative hypothesis for claims in prior work. Two independent datasets are analyzed. The analyses with the most and least predictive words are clever, and nicely complement the more naturalistic analyses.*
>
> *Weaknesses:*
>
> *I have to admit that I have a hard time understanding one conceptual aspect of the work, and a few technical aspects of the analyses are unclear to me. Conceptually, I am not clear on why stimulus dependencies need to be different from those of prediction. Yes, it is true that actively predicting an upcoming word is different from just letting the regression model pick up on stimulus dependencies, but given that humans are statistical learners, we also just pick up on stimulus dependencies, and is that different from prediction? Isn't that in some way, the definition of prediction (sensitivity to stimulus dependencies, and anticipating the most likely upcoming input(s))?*

We thank the reviewer for this comment, which highlights that the previous version wasn't sufficiently clear. Conceptually, the difference is critical: it is the difference between passively encoding or representing the stimulus (like e.g., a spectrogram of the stimulus would), and actively generating predictions.

We have substantially changed the framing of the paper to put the notion of control systems centre-stage. One such control system is the speech acoustics: they encode the stimulus (and thus its dependencies) but cannot predict. When we observe the "hallmarks of prediction" in acoustics, this demonstrates the hallmarks can arise without any prediction.

> *This brings me to some of the technical points: If the encoding regression model is learning one set of regression weights, how can those reflect stimulus dependencies (or am I misunderstanding which weights are learned)? Would it help to fit regression models on for instance, every second word or something (that should get rid of stimulus dependencies, but still allow to test whether the model predicts brain activity associated with words)? Or does that miss the point? I am a bit unclear as to what the actual "problem" with the encoding model analyses is, and how the stimulus dependency bias would be evident. It would be very helpful if the authors could spell out, more explicitly, the precise predictions of how the bias would be present in the encoding model.*

Different weights are estimated per time point in the time-resolved regression. This allows the model to learn how the response to words unfolds, but also to learn different stimulus dependencies at each timepoint. Fitting on every second word would reduce but not eliminate the problem. Our control system approach provides a more principled test. We have clarified the mechanism in the Introduction (lines 82-90), explaining how correlations between neighbouring words allow the regression model to predict prior neural activity without assuming pre-activation.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *At a high level, the reviewers demonstrate that there is an explanation for pre-word-onset predictivity in neural responses that does not invoke a theory of predictive coding or processing. The paper does this by demonstrating that this predictivity can be explained solely as a property of the local mutual information statistics of natural language. That is, the reason that pre-word onset predictivity exists could simply boil down to the common prevalence of redundant bigram or skip-gram information in natural language.*
>
> *Strengths:*
>
> *The paper addresses a problem of significance and uses methods from modern NeuroAI encoding model literature to do so. The arguments, both around stimulus dependencies and the problems of residualization, are compellingly motivated and point out major holes in the reasoning behind several influential papers in the field, most notably Goldstein et al. This result, together with other papers that have pointed out other serious problems in this body of work, should provoke a reconsideration of papers from encoding model literature that have promoted predictive coding. The paper also brings to the forefront issues in extremely common methods like residualization that are good to raise for those who might be tempted to use or interpret these methods incorrectly.*
>
> *Weaknesses:*
>
> *The authors don't completely settle the problem of whether pre-word onset predictivity is entirely explainable by stimulus dependencies, instead opting to show why naive*

> *attempts at resolving this problem (like residualization) don't work. The paper could certainly be better if the authors had managed to fully punch a hole in this.*

We thank the reviewer for their assessment.

We believe our paper does punch the hole that can be punched, which is a hole in the method. Our control demonstrates that adjusting the features (X matrix) cannot address dependencies that persist in the signal itself (Y matrix). Because the hallmarks emerge in a system that cannot predict (even after linearly removing the previous stimulus) attributing pre-onset encoding performance to neural prediction (rather than stimulus structure) is fundamentally ambiguous, and different (e.g. variance partitioning) approaches would suffer from the same ambiguity. We have reframed the manuscript to make this argument more clearly.

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *The study by Schönmann et al. presents compelling analyses based on two MEG datasets, offering strong evidence that the pre-onset response observed in a highly influential study (Goldstein et al., 2022) can be attributed to stimulus dependencies, specifically, the auto-correlation in the stimuli—rather than to predictive processing in the brain. Given that both the pre-onset response and the encoding model are central to the landmark study, and that similar approaches have been adopted in several influential works, this manuscript is likely to be of high interest to the field. Overall, this study encourages more cautious interpretation of pre-onset responses in neural data, and the paper is well written and clearly structured.*
>
> *Strengths:*
>
> *(1) The authors provide clear and convincing evidence that inherent dependencies in word embeddings can lead to pre-activation of upcoming words, previously interpreted as neural predictive processing in many influential studies.*
>
> *(2) They demonstrate that dependencies across representational domains (word embeddings and acoustic features) can explain the pre-onset response, and that these effects are not eliminated by regressing out neighboring word embeddings - an approach used in prior work.*
>
> *(3) The study is based on two large MEG datasets, showing that results previously observed in ECoG data can be replicated in MEG. Moreover, the stimulus dependencies appear to be consistent across the two datasets.*

We'd like to thank the reviewer for their comments on our preprint.

> *Weaknesses:*
>
> *(1) To allow a more direct comparison with Goldstein et al., the authors could consider using their publicly available dataset.*

We thank the reviewer for this suggestion. The Goldstein dataset was not publicly available when we conducted this research. However, we have now applied our control analyses to their stimulus material, and found that the exact same problem applies to their dataset, too.

We have added analyses of the Goldstein et al. (2022) podcast stimulus throughout the paper. Results are shown in Figures S2B, S3B, S5C, and S6B. Critically, we observe the same pattern: both hallmarks emerge in the acoustic control system, and residualisation fails to eliminate

them. This demonstrates that our findings generalise to the very dataset used to establish pre-onset encoding as evidence for neural prediction.

> *(2) Goldstein et al. already addressed embedding dependencies and showed that their main results hold after regressing out the embedding dependencies. This may lessen the impact of the concerns about self-dependency raised here.*

We thank the reviewer for raising this point, as it reveals we failed to convey a central argument in the previous version. Goldstein et al.'s control analysis did not address the concern. We show that even after the control analyses that Goldstein et al. perform (removing bigrams, regressing out embedding dependencies) the "hallmarks of prediction" still emerge when applying the analysis to a passive control system that by definition does not predict: the speech acoustics. We now also show this in their data.

To better convey this critical point, around the concept of "passive control systems". We now first establish that the hallmarks appear in acoustics (Figure 3), then show that residualisation fails to remove them (Figure 4). This makes explicit that any claim about "controlling for dependencies" must be validated against a system that cannot predict.

> *(3) While this study shows that stimulus dependency can account for pre-onset responses, it remains unclear whether this fully explains them, or whether predictive processing still plays a role. The more important question is whether pre-activation remains after accounting for these confounds.*

We thank the reviewer for this question, and we agree that the question whether pre-activation occurs is an important and interesting one. However, we ask a different question in our study: Our goal is not to definitively establish whether the brain predicts during language processing; it is to scrutinise what counts as evidence for prediction, and to correct for some highly influential claims made in the literature. The reviewer asks whether pre-activation remains "after accounting for these confounds." But the point we are trying to make is that in this analytical framework, one cannot analytically account for these confounds: corrections to the X matrix leave dependencies in the data itself intact, as the acoustic control demonstrates.

We do offer recommendations for future work. The passive control systems approach can serve as a benchmark: pre-onset neural encoding (or decoding) can only count as evidence for prediction if it exceeds what is observed in a passive control system like acoustics (which is not what we observe). Additionally, the field could move toward less naturalistic stimuli with tighter experimental controls, reducing the correlations that make this attribution so difficult. Developing a new definitive test is beyond the scope of our paper, but we believe applying this benchmark is a necessary first step.

To make this clearer, we have rewritten the Discussion to explicitly state this criterion (lines 331-340) and to outline these recommendations for future work (lines 337-340). We have also added a paragraph extending our argument to decoding approaches (lines 343-354), noting that the same ambiguity applies regardless of analytical direction.

> ***Recommendations for Authors:***
>
> ***Reviewer #1 (Recommendations for Authors):***
>
> *As per my "Weakness" point, I would appreciate engagement with the conceptual point related to the difference between prediction and stimulus correlations. Most importantly, I hope the authors will spell out more explicitly which predictions their proposal makes, and how exactly those would be present in an encoding model.*

Our proposal makes a clear prediction: if pre-onset encoding can be explained by stimulus dependencies (essentially a confound in the analysis) the same hallmarks should emerge in passive control systems that encode the stimulus but do not predict. We test this with word embeddings and speech acoustics, and both show hallmarks despite not doing any prediction.

> ***Reviewer #2 (Recommendations for Authors):***
>
> *I greatly enjoyed reading the paper and only have minor quibbles. The work is overdue and will no doubt be a valuable addition to the literature to push back on over-hyped claims about the implications of pre-word predictivity in neural response. I have few issues with the methods that the paper uses, they seem sensible and in line with previous work that has investigated these questions, and I did not find typos.*
>
> *One point I would like to raise is whether or not there is a more effective solution to resolving the issues behind residualization that the paper demonstrates. The authors show that removing next-word information does not effectively resolve the problem that local relationships in the stimulus dataset pose. The challenge to me here seems to be that it is difficult to get a model to "not learn" a relationship that is learnable. I wonder if a better solution to this is to not try to get a model to exclude a set of information but instead to do some sort of variance partitioning where you train a model to predict the next-word representation from the current-word representation (as in the self-predictivity analysis) and then build an encoding model out of the predicted representation. Then, compare the pre-word-onset encoding performance of the prediction with the pre-word-onset encoding performance of the original representation. If the performance of the two models roughly matches, that would be strong evidence that most of what these models are capturing before word onset is just explainable by the stimulus dependencies, no?*

We would like to thank the reviewer for their kind words and positive appraisal!

The proposed analysis is that if a linear proxy representation, w_hat_t – predicted linearly from w_{t-1} – yields pre-onset predictivity comparable to the actual w_t vector, this would support that the effect can be explained by stimulus dependencies. While this is an interesting alternative analysis, we would be cautious about the inverse conclusion: that if w_t outperforms the linear proxy w_hat_t, the residual variance must reflect true neural prediction.

This is because of our control system results. We show that even when we remove the "predictable" shared variance – which is similar to computing the difference between w_t and w_hat_t – the unique information still yields pre-onset predictivity, albeit reduced, in the passive acoustics that by definition cannot predict. Therefore, instead of developing an ever-more-clever way to "correct" for the problem by adjusting the X matrix, we focus on showing that the problem lies in the stimulus itself. For the revision, we focused on reframing the problem and hope we have punched a fuller hole in the logic by breaking down the fundamental issue more clearly and showing it applies to the stimulus material of Goldstein et al. (2022) as well.

Additionally, I would say that I was a bit confused about what was going on in the methods figures, to the point where I do not see the value in having them, but thankfully, the text was clear enough to resolve that confusion.

We are sad the methods illustration wasn't helpful. In presentations we have found that the illustrations were generally helpful to bring the analysis across, e.g. the aspect of keeping the analysis identical but simply replacing the brain data with either word vectors (current Figure 2) and acoustics (current Figure 3). In the revision we have reorganised the schematics

slightly, we introduce the acoustics as a control system earlier, to separately introduce residualisation and its insufficiency (Figure 4). We hope this helps

> ***Reviewer #3 (Recommendations for Authors):***
>
> *(1) My major concern is the extent to which this study offers new insights beyond what was already demonstrated in Goldstein's work. First, the embedding dependency highlighted by the authors seems somewhat expected, given how these embeddings are constructed: GloVe embeddings are based on word co-occurrence statistics, and GPT embeddings are combinations of embeddings of preceding words. More importantly, Goldstein et al. addressed this issue by regressing out neighboring word embeddings. This control was effective, as also confirmed by the current manuscript, and their main results remain. Therefore, the embedding dependency appears to have been properly accounted for in the earlier study.*
>
> *Building on the previous point, I appreciate the analysis of dependencies across representational domains, which I see as the main novel contribution of this manuscript. I would encourage the authors to explore this aspect more deeply. If I understand correctly, stimulus dependencies may persist even after regressing out neighboring word embeddings due to two potential factors:*
>
> *(a) Temporal dependencies in embeddings: since the regression of neighbor words is performed at the word level rather than over time, temporal dependency may remain.*
>
> *(b) Cross-feature dependencies - specifically, correlations between embeddings and acoustic features.*
>
> *Regarding the first factor, it is not entirely clear to me whether this is a real problem—i.e., whether word-level regression fails to remove temporal dependencies. A simulation could help clarify this and support the argument. While it's not essential, it would be valuable if the authors could propose a method to address this issue, or at least outline it as a direction for future work.*
>
> *For the second point, it would be helpful for the authors to explicitly explain the potential relationship between word embeddings and acoustic features. Additionally, while correlations between features are a common problem in speech research, they are typically addressed by regressing out acoustic features early in the analysis (Gwilliams et al., 2022). It would strengthen the current findings if the authors could test whether the self-predictability persists even after controlling for neighboring embeddings and acoustic features.*

We appreciate the extensive and detailed engagement with our work, which has been very useful in highlighting key unclarities and gaps we had to address.

We do believe our study goes well beyond what was shown by Goldstein, by identifying a fundamental limitation in their analysis, and showing that their purported control analyses do not in fact control for the problem. We'll address the reviewers' sub-questions in turn.

(i) Why this offers crucial insights beyond Goldstein et al.

While Goldstein et al. indeed addressed embedding dependencies via residualization (or in their case projection), their conclusion relied on the assumption that any neural encoding surviving this "fix" must reflect genuine predictive pre-activation. Our study invalidates this assumption. By applying the residualization fix, we show that the "hallmarks of prediction" persist just as robustly in a passive control system that cannot predict (the speech acoustics) as in the neural data. (We also show this for bigram removal.)

This provides a key new insight: persistent pre-onset predictivity after “correction” is not evidence that the dependency issue was solved. Instead, because the same effect persists in a system that cannot predict (acoustics), the persistence of the hallmarks cannot be attributed to prediction. It demonstrates that the standard "fix" is mathematically insufficient to remove the confound, rendering the original evidence for neural prediction fundamentally ambiguous.

(ii) Why do dependencies/hallmarks persist after residualization?

Residualization successfully removes the linear dependency between the current embedding (w_t) and the previous embedding (w_{t-1}) within the feature space. However, it does not (and cannot) remove the dependency from language itself, and therefore from the brain which (in some format) encodes the linguistic stimulus. Language is massively redundant. Knowing the current word tells you something about what came before – acoustically, syntactically, semantically. As long as the embedding identifies the word, the regression model will re-learn this relationship. For instance, in the case of acoustics, even when using the corrected embedding, the regression will re-learn that certain words (e.g., "Holmes") tend to follow certain acoustic patterns (e.g., the acoustics of "Sherlock"). “This shows that correcting the embeddings is insufficient: the dependencies exist in language itself, and the model will re-learn them from any signal that encodes that language.”

(iii) Why not regress out the acoustics?

This is also why "regressing out acoustics" (as the reviewer suggests) would miss the point. We do not claim that acoustic features leak into the neural signal or that acoustics are a specific confound to be removed. Rather, we use acoustics as a “passive baseline”: a system that encodes the stimulus but cannot predict. That the method yields "hallmarks of prediction" in this baseline demonstrates these hallmarks are not valid evidence for prediction—regardless of what additional features one regresses out. This motivates our proposed criterion: future studies seeing evidence for neural pre-activation should not rest on finding pre-onset encoding per se, since passive systems show this too. Rather, it should require demonstrating that the brain signal contains more information about the upcoming word than the passive stimulus baseline.

As these aspects are fundamental to the interpretation of our study, we have fundamentally re-organised and re-wrote large parts of the paper. We hope it is much clearer now.

> *(2) To better compare to Goldstein's work, the author may consider performing the same analyses using their publicly available dataset.*

This is a good suggestion. When we initially conducted this research, the Goldstein dataset was not yet publicly available. It now is, and we have applied our analyses to their stimulus material. The same problem emerges: the hallmarks of prediction appear in the acoustics of their podcast stimuli. Even after applying the control analyses, pre-onset predictivity is robust in their acoustics (indeed, in correlation terms, higher than reported for the neural data, so there is not more predictivity in the brain than in the stimulus material), confirming that the issue we identify applies to the original dataset. Results are shown in Figures S2B, S3B, S5C, and S6B.

> *(3) It is also interesting to show the predictability effect after word onsets for self-predictability analyses, for example, in Figure 2C. The predictability effect is not only reflected in pre-onset responses but also in post-onset responses, i.e., larger responses for unpredicted words. Whether the stimulus dependency mirror this effect?*

Our paper focuses specifically on temporal dependencies – the capacity of the current word to predict the previous stimulus signal (e.g., previous acoustics, previous embeddings) – and

how this mimics neural pre-activation. Post-onset analyses, by contrast, concerns the mapping between the current word and its concurrent signal, which involves fundamentally different mechanisms (e.g., mapping fidelity, frequency effects, acoustic clarity, word length) and would require the consideration of covariates of the attributes of the word post-onset to meaningfully interpret. Post-onset, there can be differences between predictable and non predictable words – e.g. sometimes unpredictable words are pronounced with more emphasis – which is why surprisal studies include a large range of covariates. However, this is not about stimulus dependencies or pre-activation, so we consider it is beyond scope of our study.

> *(4) The authors might consider reporting the encoding performance for the residual word embeddings, similar to Figure S6B in Goldstein's paper. This would allow us to determine whether pre-activation persists in the MEG responses and compare its pattern with the predictability of pre-onset acoustics.*

We do report this analysis, in the revised supplement it is shown in Figure S7. We placed it in the supplement precisely because residualized embeddings are not the "fix" they appear to be: as we show, they still yield strong pre-onset predictivity in the passive acoustic baseline (Figure 4, S6), undermining their use as a control.

> *(5) The series of previous pre-activation analyses proposed fruitful findings, e.g., the difference between brain regions (Fig. S4, (Goldstein et al., 2022)) and the difference between listeners and speakers (Figure 2, (Zada et al., 2024)). Whether these observed differences can be explained by the stimulus dependency?*

We appreciate this question. Our goal is to address the general logic of using pre-onset encoding as evidence for prediction, rather than to critique every finding in specific papers, especially as it pertains to a specific author. But briefly:

Speaker vs. Listener differences (Zada et al., 2024): Zada et al. report distinct temporal profiles: speaker encoding peaks pre-onset (planning?), whereas listener encoding peaks post-onset but shows a pre-onset "ramp." Our critique applies to interpreting this ramp as "prediction." However, this interpretation is not central to their paper, which focuses on speaker-listener coupling via shared embedding spaces. We leave the implications (which are clear enough) to the reader.

Regional differences (Goldstein et al., 2022): Encoding timecourses do vary across electrodes, as we also observe across MEG sources (and participants). But our point is logical: because pre-onset encoding does not necessarily reflect prediction, finding a channel with stronger pre-onset encoding does not mean that channel performs “more prediction”. For instance, one subject in the Armeni dataset showed higher pre-onset than post-onset encoding (and indeed activity) overall – but it would be implausible to conclude this subject "only predicts" and does not “process” or “listen”. More likely, this reflects differences in signal-to-noise, integration windows, or source contributions. The exact sources of these morphological differences are interesting but unclear, and speculating on them is beyond our scope.

> *(6) I appreciate that the authors have shared their code; however, some parts appear to be missing. For example, the script encoding_analysis.py only includes package-loading code.*

Thank you for noticing, we have updated our code database.

> *(7) What do the error bars in the figures represent - for example, in Figure 1C? How many samples were included in the significance tests? The difference between the two curves appears small, yet it is reported as significant. Additionally, Figure S1 shows large differences between subjects and between the two MEG datasets. Do the authors have any explanation for these differences?*

The shaded areas in our previous Figure 1c) show 95% confidence intervals computed over the 100 MEG sources identified to be part of the bilateral language system and the 10 cross-validation splits.

We do not have an elaborate explanation for the differences in encoding performance across the three subjects in the few-subject dataset. Instead, we interpret these differences as a likely consequence of substantial inter-individual variability in evoked responses, even at the source level, arising from differences in cortical folding and the orientation of underlying current dipoles. We deem this a likely explanation since different electrodes in Goldstein's ECoG data also showed very different encoding profiles.

With respect to the multi-subject dataset, we suspect that the large differences stem most likely from two substantial differences: First, the acoustics were purposefully manipulated by the experimenters to reduce temporal dependence. This made it harder for listeners to concentrate on the stories and thereby might have potentially led to lower quality neural data. Furthermore, it reduced one form of stimulus dependency, namely the acoustic temporal dependencies, which could be exploited by the encoding model to reach higher encoding accuracies. Secondly, MEG has a notoriously poor signal-to-noise ratio, and the amount of data per participant (7.745 words as opposed to 85.719 in the few-subject dataset) might not have been enough to produce reliably high encoding results.

> *Finally, the current study is clear and convincing, and my suggestions are not intended to question its novelty or robustness. Rather, I believe the authors are in a strong position to address a critical question in language processing: whether pre-activation occurs. The authors have thoughtfully considered important confounds related to pre-onset responses. Adding some approaches to regressing out these confounds could be particularly helpful for determining whether a true pre-onset response remains.*

We thank the reviewer again for their constructive feedback, suggestions and questions. To clarify, however, our goal is *not* to definitively attest to whether pre-activation occurs. Our goal is simply to scrutinise a specific method to test for linguistic prediction. This method purports to be an improvement on conventional post-onset (e.g. surprisal-based) methods, as it can directly investigate effects occurring prior to word onset. We have demonstrated fundamental limitations in the underlying logic of this method. We propose passive control systems as baselines against which claims of prediction should be evaluated. Against this baseline, the current evidence does not show unequivocal support for prediction: pre-onset encoding in the brain does not exceed that in the passive control. However, we do not conclude from this that pre-activation does not exist — that would require a different study entirely. Our aim is more methodological: to establish what should count as evidence for prediction, not to settle whether prediction occurs.

We would like to thank the reviewers and editors for their thoughtful feedback, which has been tremendously helpful in improving the paper.

https://doi.org/10.7554/eLife.106543.2.sa0