nature research

Peer Review Information

Journal: Nature Computational Science
Manuscript Title: Combining computational controls with natural text reveals aspects of meaning composition
Corresponding author name(s): Leila Wehbe

Reviewer Comments & Decisions:

Decision Letter, initial version:

Date: 21st January 22 13:51:10
Last Sent: 21st January 22 13:51:10
Triggered By: Kaitlin McCardle
From: kaitlin.mccardle@us.nature.com
To: lwehbe@cmu.edu
BCC: kaitlin.mccardle@us.nature.com
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-21-0986
Message: ** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Dr Wehbe,

Your manuscript "Combining computational controls with natural text reveals new aspects of meaning composition" has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address all of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Please provide quantitative comparisons to other architectures, including unidirectional language models
- Please demonstrate what *specifically* the residual embeddings represent (i.e. what

is the "added value" in the residual embeddigns from the LSTM?)
- Please justify the use of LSTM versus a BERT approach
- If large datasets are not publicly available to perform experiments on, please explain this in your response to referees' letter and justify why your current dataset is adequate.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making of use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best regards,

Kaitlin McCardle
Editor
Nature Computational Science

Reviewers comments:

Reviewer #1 (Remarks to the Author):

Review of manuscript „Combining computational controls with natural text reveals new aspects of meaning composition"

This is an interesting study investigating the neural correlates of what the authors call "supra word meaning" – the combined meaning of words beyond their individual meaning. The authors devise a computational measure for supra word meaning and predict fMRI and MEG data from this measure. They find interesting correlates of supra word meaning in fMRI suggesting a shared neural basis for word level and composed meaning. On the other hand, their measure of supra word meaning does not predict MEG activity.

I found the general approach and the results obtained from the fMRI data very interesting. The fMRI results seem convincing as they are replicated across two very distinct datasets obtained during reading versus watching a movie. The results are theoretically relevant and they seem to nicely converge with the assumptions implemented in a neural network model of sentence processing, which does not assume separate steps of lexical retrieval and integration during sentence comprehension (Rabovsky, Hansen, & McClelland, 2018, Nature Human Behavior).

My main problem with this paper is that far reaching implications concerning the brain basis of meaning composition (namely "that composed meaning is maintained through a different neural mechanism than synchronized firing of pyramidal cells"; abstract) are drawn based on what is basically a null result – namely the absence of an influence of the specific measure of supra word meaning devised by the authors on MEG activity recorded from only 8 participants.

Before drawing such far reaching conclusions (see also parts of the discussion p. 13-16), and given in general quite some disagreements in the literature even between datasets obtained using the same imaging modality, I think it would be important to replicate these MEG results in a substantially larger sample of participants. It would also be informative to get something like a power analysis. Alternatively, I would think that maybe it could also be an option to focus more on the fMRI results, and then just report that these results could not be replicated with MEG, without making such a big and fundamental issue out of this null result in the MEG data

Other comments:

p. 6: "the full context embedding is predictive of much of the fMRI recordings across the brain" – I can't help but find this problematic as it seems highly unlikely that much of the brain is devoted to meaning processing (or whatever processes are thought to be indexed by the full context embedding). I am aware that this is not the focus of the authors work, but because the authors' general approach is based on this type of prediction based analysis of neural data I was wondering whether they could comment on how they interpret this result and what it might mean for the type of analysis they are using.

p. 14: "One explanation may be that because the MEG signal is thought to be most related to synchronized firing of pyramidal cells, it best reflects the representation of working memory, which is thought to be supported by pyramidal cells." Here it seems relevant that the N400 ERP component, which is thought to reflect aspects of meaning composition and is measured in EEG (and thus also depends on synchronized firing of

pyramidal cells), seems to reflect a relatively automatic process of meaning composition that takes place even during the attentional blink (Luck, Vogel & Shapiro, 1996, Nature).

Reviewer #2 (Remarks to the Author):

When words are combined into a sentence, what humans understand is, in several ways, more than the simple sum of its components. For example, the combination of a noun and a modifier lead to subtle inferences about features of the combined concept (studied in concept combination), and for expressions like "Mary finished the book", humans infer some activity that Mary did with the book, like reading and writing. This paper asks where these kinds of "added meaning", here termed supra-word meaning, can be located in the brains of human participants as they read or listen to natural discourse.

This paper takes a computational, data-driven approach to formalizing this supra-word meaning, building on recent advances in natural language processing (NLP). There, it has been shown that language models, computational models trained to predict upcoming words, or missing words in context, learn a lot about linguistic structure, word meaning, and even commonsense knowledge.

The paper makes the following interesting and novel contributions:
* It defines a new model of supra-word meaning, based on the recent ELMo contextualized language model. This model represents supra-word meaning through an embedding (a vector).
* It shows that it is possible to use these embeddings to predict fMRI voxels in particular brain regions of participants reading a story or watching a movie
* It shows that, intriguingly, it cannot predict MEG responses

All these contributions are interesting and novel. However, there is a big gap in this paper that I think needs to be fixed first. This gap concerns the ELMo-based representations of supra-word meaning.

The paper lists several kinds of meaning that could, in principle, form part of this supra-word meaning. This includes coarse-grained word sense disambiguation, such as "play a game" versus "theater play"; fine-grained context effects on word meaning (possibly below the level of stored senses), such as "green banana" being understood as unripe; coercion/regular polysemy effects such as "Mary finished the book" being understood as "finished reading" or "finished writing"; and differences in semantic role assignment ("Mary kicked John" versus "John kicked Mary"), which could also include semantic proto-role inferences (though the authors don't say this), such as: if "Mary kicked John" then Mary probably volitionally initiated the event, while John was a nonvolitional participant who may have been changed by the event.

Anyway, my point with this is that this is a long list of possible meaning-in-context effects studied in the linguistic and psychological literature -- but which of those, if any, are encoded in the ELMo-based residual embeddings that the authors define? This is very important to know, because we can only draw conclusions from the connection between these embeddings and fMRI activation if we know what knowledge the

embeddings encode.

For many standard embeddings, we have a good sense of what knowledge they encode, because there are tons of paper on probing those embeddings, or predicting linguistic features or human participant reactions from those embeddings. But the residual embeddings defined here are new, and we do not know. So I think what needs to be done first is to probe these new residual embeddings so we can know what they are a model of: Can they be used to predict the original word, or one of its senses? Can they be used to predict sentence similarity, or reversals in semantic role assignment? Or can we deduce, from prior probing work on ELMo or LSTMs in general, what the embeddings must contain? I am sorry -- I know this is a lot to ask, but I feel that it is necessary.

To compute residual embeddings, the authors train simple linear regression models with ridge regularizers to predict an ELMo hidden layer embedding (contextualized embedding) from a sequence of static embeddings for the input words. (Plus two other residual computations, which predict input embeddings from other input embeddings plus a hidden-layer embedding.) So the residual, the proportion of the contextualized embedding not predicted by the linear function, should in principle contain the "added value" from the LSTM, what it has learned about ignoring versus integrating inputs, from the task of predicting the next word of the sequence. It is not clear that all of the kinds of semantic information listed above would be relevant to the language modeling task, so some of them may not be represented in the residual embedding that much.

--
Other comments:

Introduction: The picture of linguistic research on sentence meaning is quite simplified. The text implies that linguistics only considers "semantic structure" and has nothing to say about word meaning, but for example Nicholas Asher's book on lexical meaning, or the radical contextualist view of meaning in context, do have a lot to say about, for example, the difference of 'red apple' and 'red watermelon', or interpretations of 'green leaves' (Travis 1997). Also I didn't quite understand why the paper cites this particular Partee paper, and only this one. The paper is about noun phrase typing -- is this supposed to exemplify linguistic approaches to typing?

Embeddings: The paper uses ELMo embeddings, which are recent but not quite that recent, and which in NLP have been superseded by Transformer-based models such as BERT, RoBERTa, Electra etc. I agree that it makes sense to use an LSTM-based model for modeling neural responses, and not such a gigantic black box as BERT and friends, but maybe add a sentence or a footnote about this choice.

What data was the ELMo model pre-trained on?

Reviewer #3 (Remarks to the Author):

This paper proposes to utilize a data-driven computational approach to study the brain

representation of the combined meaning of words beyond their individual meaning which is termed as "supra-word meaning". The key to this computational approach is to calculate the embeddings of "supra-word meaning" which is modelled by the residual information in the context embeddings after removing information related to the word embeddings, in which both word embeddings and context embeddings are from a pretrained language model ELMo. The ideas sound reasonable. However, there are some problems as follows:

1. ELMo is a bidirectional language model with a forward LSTM and backward LSTM.This paper only used results from forward LSTM, but the backward LSTM could have influence on the word and contextual embeddings during training. I think it is more reasonable to increase an experiment using unidirectional language model such as GPT2, and detailed comparison of experiments using different models is necessitated. It is expected to see will different language model architecture have the different influence on the brain encoding results?

2.A related question is does the embedding calculated by context(t-1)-g(word(t), word(t-1)) really capture the "supra-word meaning" information? This hypothesis has never been testified in the paper. This is very important and should be testified before the experiments of brain encoding. If it doesn't, then other experiment will be the buildings in air. It is meaningless. There are some ways to testify this such as to build a test dataset including phrases with "supra-word meaning" along with some human annotations. Moreover, in the introduction, the "supra-word meaning" is defined as the composed meaning of a sequence of words that is not part of the corresponding bag-of-words, i.e., implied meaning and others. However, it is not clearly illustrating whether there are or how many such sentences with "supra-word meaning" in the neuroimaging stimuli in experiments? In the paper, Chapter 9 of Harry Potter novels is taken as neuroimaging stimuli, but it doesn't say what proportions of this type of sentences are contained in Chapter 9? If it's not the majority, then the residual embedding cannot be taken as that encoded the "supra-word meaning".

3.There are also some technical details that are not clearly described. Why only the first hidden layer is utilized as the contextual embedding? The second hidden layer also encodes the contextual information. Will the results be different if the second hidden layer or the average of the two layers are utilized? Why a ridge regression method is employed to calculate the linear function g? Will there be difference using other method such as one with an analytical solution?

4.After obtaining the "supra-word meaning" representations, this paper builds the brain encoding models on two fMRI datasets and one MEG dataset. For the brain encoding method, the following technical details are not clearly explained.
1)To align brain signal with embeddings, this paper chooses to use concatenate vectors of several time steps to model the delay, but the more conventional way is to use HRF. Why authors don't utilize HRF? Are there any special reasons or considerations?

2)For the cross-validation procedure, how to choose the training and validation data? If the training and validation data are too close in time, then their neuroimaging data should be very similar, which will lead to an overfitting result.

3)In the permutation test, why you choose 5TRs as a block?

4)For the brain encoding results, different subjects show the different patterns especially for the residual context embeddings. Moreover, the anterior and posterior temporal lobes are broad. I am not sure this is enough to verify that "supra-word meaning is predictive of ATL and PTL". How to explain the difference between subjects? Is there a quantitative indicator to calculate a significant group-level results?

5)The results from Harry Potter fMRI and movie fMRI are very different, and there is not any quantitative indicator to explain their similarities and differences. Is it possible to use the method in Figure 4 to calculate the relation between two fMRI dataset? To compare the two results, it is better to give the significant-voxel results (in Figure S1) for the movie data.

6)The findings of "supra-word meaning is not detectable in MEG" are unexpected and may needs more experiments to verify. There are several factors that could affect the results. For instance, the serial visual format is not a natural way to understand language, and the results may be different when collecting MEG data using a natural audio listening paradigm. In addition, there are only 8 subjects for the MEG experiments and the individual results are not shown in the paper. Is there any subject having significant results for the supra-word embedding? When preprocessing the MEG data, does the time delay between the stimulus and the recording of the actual data be considered?

BTW, there is a minor typos error, i.e., Figure1 lacks title D.

In summary, I think there are much room to improve the manuscript, and the work can be made more solid.

Author Rebuttal to Initial comments

Author Response: Combining computational controls with natural text reveals new aspects of meaning composition

We thank the reviewers and editors for their time and valuable feedback. We believe that addressing the reviewers' concerns strengthened the manuscript. In response to the reviewers, we include four new analyses in the edited manuscript, which we summarize in Section 1. We further respond to several important concerns that are shared between reviewers in Section 2. We lastly include point-by-point responses to the reviewers in Section 3 with references to changes in the manuscript. Line numbers cited here correspond to lines in the version of the manuscript labeled with changes.

1 Summary of new analyses

The edited manuscript contains four new analyses:

1. A replication of the supra-word vs full context results in the fMRI Harry Potter recordings using a second NLP model (GPT-2).
    - This analysis replicated the results using ELMo, showing that all bilateral language ROI are predicted significantly by the full context representations, and that the bilateral ATL and PTL are predicted significantly by the supra-word meaning (i.e. residual context representation). These results are presented in Figure R3 in the response document and in Suppl. Fig. S5.
2. A replication of the same analysis in MEG.
    - This analysis replicated the results using ELMo, showing that while the full context representation predict the MEG recordings significantly, the supra-word meaning representations fail to significantly predict any sensortimepoint across participants. These results are presented in Figure R4 in the response document and in Suppl. Fig. S14.
3. An analysis of the difference between supra-word and full context representations.
    - We show that the full context representation is most (linearly) related to the current word, the previous two words, and the future word. In contrast, we show that the supra-word meaning indeed does not strongly relate to any of these words, as it was designed. These results are presented in Figure R2 in the response document and in Suppl. Fig. S1. We use this analysis to shed some light on the content of the supra-word representations.
4. An analysis with pilot data for an MEG natural speech listening experiment.

- This analysis replicates the results from the MEG reading data. Notably, while the current word and the context embeddings are predictive of activity, the supra-word meaning is not.

We expand on these new analyses and results in the following sections.

2 Clarification on shared concerns

In light of the null result that we are not able to predict any MEG sensor-timepoint with the supra-word meaning representation, several reviewers voiced concerns over the quality of our MEG data and the quality of the supra-word meaning representations.

We would like to point out that while the supra-word meaning vector cannot predict MEG data, it can predict the fMRI data, and the MEG data is well predicted by other vectors, and thus we believe that the quality of the MEG data and the supra-word meaning embedding is good:

- *We emphasize that the null result is accompanied by two strongly significant results which, when taken together, should alleviate some concerns about the quality of the MEG data and the quality of the supra-word meaning embeddings. These two strongly significant results are that: 1) the MEG data is indeed strongly predicted by representations of both individual words and context (Fig. 3 in the main paper), and 2) the supra-word meaning embedding predicts significant proportions of the fMRI recordings (Fig. 1 in the main paper). There is a null result* only *when the supra-word meaning embeddings are used to predict the MEG data. (We also want to highlight that for reproducibility purposes, there is value in reporting negative results in papers and not file-drawering them.)*

- *Additionally, all individual participants show nonexistent to extremely low supra-word meaning prediction in MEG (Response Fig. R7): Of the 6120 total MEG sensor-timepoints (306 sensors x 20 timepoints) per subject, supra-word meaning predicts the following number significantly (at 0.05 level, permutation test followed by FDR correction): 0, 0, 1, 2, 5, 5, 20, 21. Note that these sensor-timepoints do not co-occur across participants. In contrast, the context predicts the following number of sensor-timepoints significantly: 3050, 4203, 3963, 1636, 3010, 5100, 4714, 4702. We provide the individual-level helmet plots for surpaword meaning in Fig. R6 in the response document and in Suppl. Fig. S13. Given these results, we are confident that removing the individual word information from the context eliminates much of the information that is useful to predict the MEG recordings.*

Relatedly, the reviewers asked for a replication of the analyses using a larger MEG dataset.

We are not aware of a publicly available MEG dataset with more subjects and a similarly large number of trials-per-subject.

- *Both the number of subjects and the numbers of trials per subject are important when considering the significance of the results[1].*

- *The MEG dataset that we used has 5000 samples per subject and 8 subjects.*

- *This is enough data for the strongly significant results of predicting the MEG data with representations of individual words and with the full context (Fig. 3 in the main text), both at the individual subject and the aggregate group level.*

- *We would be happy to replicate our supra-word meaning experiments with another MEG dataset that has potentially greater power (more than 8 subjects and a large number of trials per subject). We are not aware of such a dataset that is publicly available. The closest dataset that we are aware of is the MOUS dataset[2]. This dataset has 102 subjects but only 120 trials per subject. Furthermore, the stimuli are in Dutch, which would require using a different NLP model that was trained using Dutch. If the reviewers or editors can point us to an openly available MEG dataset with more than 8 subjects and a large number of trials per subject, we would be thankful and be very happy to replicate our experiments.*

- *Coincidentally, we had acquired pilot data for one subject listening to 70 minutes (15030 words) of stories from the moth Radio Hour (this is a replication of the Huth et al. (2016)[1] dataset in MEG, and more specifically of the first of the two sessions). While this is not an ideal dataset, it has enough datapoints that we are able to train and test encoding models with sufficient reliability for that subject. We show that our MEG results replicate in that dataset (see Figure R1). Namely, we find that the full context embedding from ELMo significantly predicts a large number of sensor/time-points (2500 of the 6120), with the peak of the prediction being around 125-350ms. The areas that are well predicted are bilateral fronto-temporal cortex, which makes sense since this is an auditory task that recruits the auditory cortex and the language regions, and has no visual component. When we use the supra-word meaning (the context residuals), we find that no sensor/time-point at all is predicted significantly (permutation test, FDR corrected for multiple comparisons). This replicates our negative results from the reading experiment. These results are included in Supplementary Figure S15 and referenced in the results (lines 171-174) and the discussion (lines 284-289).*

3 Point-by-point responses to reviewers

3.1 Reviewer 1

My main problem with this paper is that far reaching implications concerning the brain basis of meaning composition (namely "that composed meaning is maintained through a different neural mechanism than synchronized firing of pyramidal cells"; abstract) are drawn based on what is basically a null result – namely the absence of an influence of the specific measure of supra word meaning devised by the authors on MEG activity recorded from only 8 participants.

Regarding the hypothesis "that composed meaning is maintained through a different neural mechanism than synchronized firing of pyramidal cells", we have tried to propose this as just a hypothesis that derives from our results. To keep the tone moderate, we have stated that the involvement of other mechansisms than pyramidal cell firing is a hypothesis in the abstract.

Before drawing such far reaching conclusions (see also parts of the discussion p. 13-16), and given in general quite some disagreements in the literature even between datasets obtained using

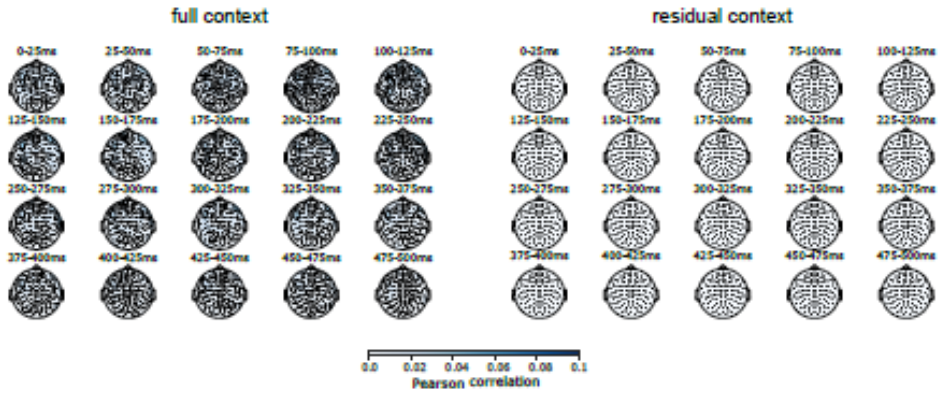

Fig. R1. Prediction performance on held-out natural speech listening data when using different delays from the onset of a word. Only the significantly predicted sensor-timepoints are displayed. Since different words have different lengths, we build a finite impulse response model that estimates the effect of a feature at each delay, following the fMRI approach. However, we only consider one delay at a time in the model (i.e. we take the design matrix containing the embeddings of the word occurring at each time step, and delay it by the corresponding amount). This allows us to have results at a single delay, like the reading MEG results. [Left] Performance using the full context feature space. Many sensor/time-points, especially starting from 125ms to 350ms, are predicted with significantly higher than chance performance. [Right] No sensor/time-point is predicted significantly by the residual context vector, replicating our negative results in the reading MEG data.

the same imaging modality, I think it would be important to replicate these MEG results in a substantially larger sample of participants. It would also be informative to get something like a power analysis.

We would like to state that the negative results with supra-word meaning don't exactly contradict existing results in the literature, because other studies have not explicitly tried to isolate supra-word meaning. These negative results suggest a hypothesis that could lead to a different interpretation of a specific aspect of what we know from existing studies. We have added a section to the discussion (lines 284-289) that highlights the importance of replicating our results with other datasets with multiple subjects (and importantly, a lot of naturalistic data per subject).

Please see Section 2 regarding replication in larger MEG datasets. We did not find an available large dataset that has enough data per participant to be able to train encoding models. Using pilot data for one additional subject with a larger number of samples than the MEG reading paradigm, we were able to replicate the negative result in MEG. Unfortunately it is not easy to obtain naturalistic reading MEG data. Regarding the number of participants, encoding model analyses routinely use a comparable number of participants. The experiments are geared towards collecting more data from fewer participants to be able to estimate reliable models for each participant. The statistics of the encoding model are complex because the modeling is done by subject (where N corresponds to the time points, which are many) and then grouped together across subjects.

Alternatively, I would think that maybe it could also be an option to focus more on the fMRI results, and then just report that these results could not be replicated with MEG, without making such a big and fundamental issue out of this null result in the MEG data.

We take the point of the reviewer and have reduced the emphasis on the interpretation of the

results, and have emphasized the necessity of replication in other datasets (see lines 283-289). However, we do think it is important to report and discuss moderately the negative results. These results are not contradictory of other published results because no other studies have done such specific controlled analyses. We believe these results and the possible hypotheses we can derive from them can lead to a new understanding of what MEG can tell us. Since this type of modeling is becoming more and more popular, we think the field would benefit from knowing our negative results.

p. 6: "the full context embedding is predictive of much of the fMRI recordings across the brain" – I can't help but find this problematic as it seems highly unlikely that much of the brain is devoted to meaning processing (or whatever processes are thought to be indexed by the full context embedding). I am aware that this is not the focus of the authors work, but because the authors' general approach is based on this type of prediction based analysis of neural data I was wondering whether they could comment on how they interpret this result and what it might mean for the type of analysis they are using.

Our wording ("across the brain") might have been a bit imprecise. We have now changed it to "much of the fMRI recordings in the language and semantic networks" on lines 115. The ROI that we consider are thought to support language comprehension and word meaning. These bilateral ROI make up 30% of the cortical voxels on average across participants. These areas have also been reported to be predicted by language meaning in other studies, such as Huth et al. (2016)[9]. These ROI are the main regions that are predicted well by the full context. We agree that the ability to predict all of these regions using the full context does not mean that they're processing the exactly same thing. The full context vector is a complex representation that contains multiple types of information (e.g. part of speech, dependency, word length, etc.) and the different ROI may be predicted well by different types of information in the full context. This can also be the case for the supra-word meaning representation, and we try to address this and understand whether the bilateral ATL and PTL process the same type of supra-word meaning by testing whether encoding models trained to predict voxels in one regions can generalize to voxels in another region (see Fig. 2C-D in the main paper).

p. 14: "One explanation may be that because the MEG signal is thought to be most related to synchronized firing of pyramidal cells, it best reflects the representation of working memory, which is thought to be supported by pyramidal cells." Here it seems relevant that the N400 ERP component, which is thought to reflect aspects of meaning composition and is measured in EEG (and thus also depends on synchronized firing of pyramidal cells), seems to reflect a relatively automatic process of meaning composition that takes place even during the attentional blink (Luck, Vogel & Shapiro, 1996, Nature).

We thank the reviewer for this very relevant point and paper and we have now included it in our discussion in lines 269-277: "One explanation may be that because the MEG signal is thought to be most related to synchronized firing of pyramidal cells, it best reflects language computations that are thought to be supported by pyramidal cells. For example, the N400 and other ERPs related to composition operations, are measurable in both MEG and EEG, and thus are likely to recruit pyramidal cells. Working memory processes have also been shown to involve pyramidal cell firing[5]. While both the N400 and working memory are thought to reflect short-term transient processing[6,7], maintaining the supra-word meaning may depend on longer-term memory processes, which may be

supported by different cell types or mechanisms.".

3.2 Reviewer 2

The paper lists several kinds of meaning that could, in principle, form part of this supra-word meaning. This includes coarse-grained word sense disambiguation, such as "play a game" versus "theater play"; fine-grained context effects on word meaning (possibly below the level of stored senses), such as "green banana" being understood as unripe; coercion/regular polysemy effects such as "Mary finished the book" being understood as "finished reading" or "finished writing"; and differences in semantic role assignment ("Mary kicked John" versus "John kicked Mary"), which could also include semantic proto-role inferences (though the authors don't say this), such as: if "Mary kicked John" then Mary probably volitionally initiated the event, while John was a nonvolitional participant who may have been changed by the event.

Anyway, my point with this is that this is a long list of possible meaning-in-context effects studied in the linguistic and psychological literature – but which of those, if any, are encoded in the ELMo-based residual embeddings that the authors define? This is very important to know, because we can only draw conclusions from the connection between these embeddings and fMRI activation if we know what knowledge the embeddings encode.

For many standard embeddings, we have a good sense of what knowledge they encode, because there are tons of paper on probing those embeddings, or predicting linguistic features or human participant reactions from those embeddings. But the residual embeddings defined here are new, and we do not know. So I think what needs to be done first is to probe these new residual embeddings so we can know what they are a model of: Can they be used to predict the original word, or one of its senses? Can they be used to predict sentence similarity, or reversals in semantic role assignment? Or can we deduce, from prior probing work on ELMo or LSTMs in general, what the embeddings must contain? I am sorry – I know this is a lot to ask, but I feel that it is necessary.

To compute residual embeddings, the authors train simple linear regression models with ridge regularizers to predict an ELMo hidden layer embedding (contextualized embedding) from a sequence of static embeddings for the input words. (Plus two other residual computations, which predict input embeddings from other input embeddings plus a hidden-layer embedding.) So the residual, the proportion of the contextualized embedding not predicted by the linear function, should in principle contain the "added value" from the LSTM, what it has learned about ignoring versus integrating inputs, from the task of predicting the next word of the sequence. It is not clear that all of the kinds of semantic information listed above would be relevant to the language modeling task, so some of them may not be represented in the residual embedding that much.

ELMo embeddings contain slowly evolving meaning because ELMo is able to contain information across a long time. However, a lot of this meaning is related to the recent words because ELMo needs to predict the next word so local information is needed. To show this empirically, we analyze the amount of weight that is put on each word in the local context when their token-level word embeddings are jointly used to linearly predict the full context representation. We present the results in Figure R2 in the response document and in Suppl. Fig. S1 in the edited paper. This result shows that the full context embedding from the first hidden layer of ELMo evaluated at word t is strongly linearly related to the next word t+1, current word t, and the previous word t-1. The presence of the strong linear relationship between the full context representation and the closely adjacent words means that any linear prediction of brain recordings may be overwhelmingly related

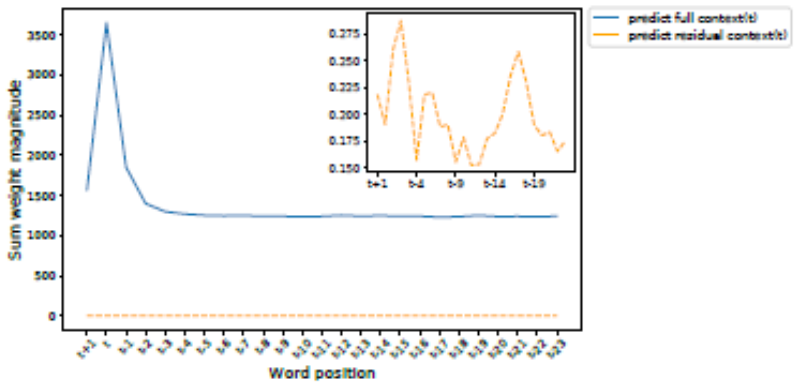

Fig. R2. Importance of individual word embeddings for predicting the full and residual context embeddings. (Blue) The full context embedding from the first hidden layer of ELMo evaluated at word t is strongly linearly related to the next word t+1, current word t, and the previous word t-1. The presence of the strong linear relationship between the full context representation and the closely adjacent words means that any linear prediction of brain recordings may be overwhelmingly related to these adjacent words. In contrast, the supra-word embedding for word t (orange) does not strongly depend on any individual word, as it was designed to limit contributions of individual words.

to these adjacent words. In contrast, we show that predicting the supra-word embedding for word t does not strongly depend on any individual word, as it was designed to limit contributions of individual words. The residual that remains corresponds to some meaning that has been non-linearly accrued over the different words. This could be considered a proxy for composition of words that leads to new meaning. The fact that it this residual is predictive of fMRI recordings means that it is not insignificant. We include this discussion in lines 305-312.

We agree with Reviewer 2 that understanding exactly what linguistic information is captured by this residual is an important research question. Understanding what linguistic information is encoded by the non-linear transformations in deep NLP models would be on its own an impactful contribution and would involve a large effort akin to the rich literature on interpreting linguistic information in NLP model representations. In the current work, we focus on showing what is not *in the supra-word representation and that this missing information is* necessary *for predicting MEG recordings well. We believe this to be an important contribution, especially with the increased recent interest in aligning representations from NLP models with brain recordings[8–10].*

We additionally point out that knowing what linguistic information is *captured by the supra-word embedding or any other stimulus representation that contains multiple types of information does not necessarily mean that this specific type of information is useful for predicting brain recordings[11]. Additional research will be needed to exactly map the specific type of information in a complex stimulus representation to its corresponding brain predictions. We include this discussion in lines 312-318.*

Introduction: The picture of linguistic research on sentence meaning is quite simplified. The text implies that linguistics only considers "semantic structure" and has nothing to say about word meaning, but for example Nicholas Asher's book on lexical meaning, or the radical contextualist

view of meaning in context, do have a lot to say about, for example, the difference of 'red apple' and 'red watermelon', or interpretations of 'green leaves' (Travis 1997). Also I didn't quite understand why the paper cites this particular Partee paper, and only this one. The paper is about noun phrase typing – is this supposed to exemplify linguistic approaches to typing?

We refer to[12] as evidence that syntactic composition is not necessarily the same as logico-semantic composition (the paper details several different possibilities for typing the same noun phrase). We now specify this more clearly in line 21. Our main goal in this paragraph is to convey that we, and others[1,3], are not aware of linguistic theories that can explain some effects observed in neurolinguistics during word composition–for instance that the LATL shows the same repeatable composition effect for adjective-noun pairs such as "red vegetable" and "tomato dish", but not for "red tomato" and "vegetable dish". For the purposes of the brief summary of the literature, we think of logico-semantic composition as encompassing the modification of lexical meaning. We now cite in line 20 the additional works suggested by the reviewer in order to clarify this.

Embeddings: The paper uses ELMo embeddings, which are recent but not quite that recent, and which in NLP have been superseded by Transformer-based models such as BERT, RoBERTa, Electra etc. I agree that it makes sense to use an LSTM-based model for modeling neural responses, and not such a gigantic black box as BERT and friends, but maybe add a sentence or a footnote about this choice.

We thank the reviewer for the suggestion and we have now added our reasoning for using ELMo in lines 77-82. In response to Reviewer 3, we also added analyses using GPT-2 which replicate our findings from ELMo (Fig. R3 and Fig. R4 and corresponding text in the response document; Suppl. Fig. S5 and Suppl. Fig. S14 in the main paper).

What data was the ELMo model pre-trained on?

The ELMo model that we used was pretrained on the 1 Billion Word Benchmark[14], which contains approximately 800 million tokens of news crawl data from WMT 2011. GPT-2 was pretrained on 8 million web pages, which were scraped using Reddit.

3.3 Reviewer 3

1. ELMo is a bidirectional language model with a forward LSTM and backward LSTM. This paper only used results from forward LSTM, but the backward LSTM could have influence on the word and contextual embeddings during training. I think it is more reasonable to increase an experiment using unidirectional language model such as GPT2, and detailed comparison of experiments using different models is necessitated. It is expected to see will different language model architecture have the different influence on the brain encoding results?

We thank the reviewer for the suggestion. We have now added results using full context embedding and a supra-word meaning embedding(e.g. residual context) obtained using GPT-2 for both fMRI (Suppl. Fig. S5) and MEG (Suppl. Fig. S14). We also refer to these figures in lines 117-121 and lines 166-171. We include these figures below for ease of discussion. We also want to clarify that in the ELMo implementation we use, the forward and backward LSTMs are trained separately, with the backward LSTM not influencing the context of the forward LSTM[15]. We emphasize this now in lines 77-78.

The fMRI results presented below in Fig. R3 replicate the results using ELMo, showing that all

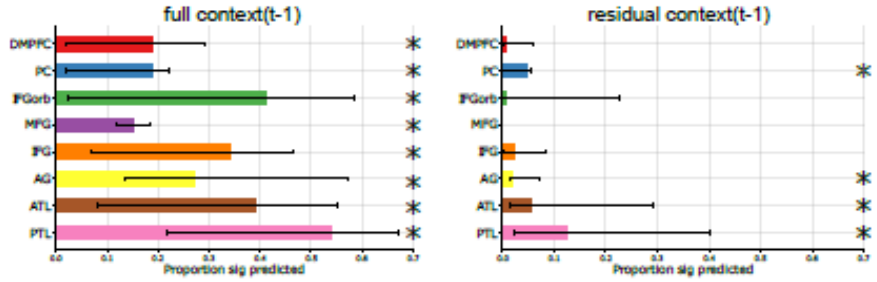

Fig. R3. fMRI prediction results for embeddings obtained using GPT-2. Proportion of language ROI voxels significantly predicted by (Left) full context and (Right) residual context embeddings. Displayed are the median proportions across all 9 participants and the medians' 95% confidence intervals. Similarly to the full context obtained using ELMo (Fig. 2B), the full context extracted from GPT-2 predicts all ROIs (ROI-level Holm-Bonferroni correction, $p < 0.05$). The residual context predicts bilateral ATL, PTL, Angular Gyrus, and Posterior Cingulate, which replicates the results using ELMo that supra-word meaning is predictive of the bilateral ATL and PTL (Fig. 2B).

bilateral language ROI are predicted significantly by the full context representations. Furthermore, they confirm that the bilateral ATL and PTL are predicted significantly by the supra-word meaning (i.e. residual context representation). In addition to the bilateral ATL and PTL, the supra-word meaning obtained from GPT-2 significantly predicts the bilateral Angular Gyrus (AG) and Posterior Cingulate (PC) across subjects. This added predictive power may be due to several factors that we cannot control without training a model from scratch: GPT-2 has larger embeddings (768 vs 512 for the forward LSTM in ELMo), has more hidden layers (12 vs 2 in ELMo), was pretrained on more data (40GB of text data vs 11GB of text data for ELMo), and has an all-together different architecture (Transformer-based vs recurrence-based for ELMo). Overall, we agree that evaluating the ability of different architectures to encode supra-word meaning is an interesting question, but it also needs to be approached with care because of these differences between the models and we believe this is a great topic for future work. We include this discussion in lines 322-330.

The MEG results presented in Fig. R4 replicate the results using ELMo, showing that while the full context representation predict the MEG recordings significantly (A), the supra-word meaning representations fail to significantly predict any sensortimepoint across participants (B). Furthermore, we see that even though the supra-word meaning extracted from GPT-2 has added predictive power over the one in ELMo for the fMRI recordings, it still fails to predict any MEG sensortimepoint significantly across participants.

2.A related question is does the embedding calculated by context(t-1)-g(word(t), word(t-1)) really capture the "supra-word meaning" information? This hypothesis has never been testified in the paper. This is very important and should be testified before the experiments of brain encoding. If it doesn't, then other experiment will be the buildings in air. It is meaningless. There are some ways to testify this such as to build a test dataset including phrases with "supra-word meaning" along with some human annotations. Moreover, in the introduction, the "supra-word meaning" is defined as the composed meaning of a sequence of words that is not part of the corresponding bag-of-words, i.e., implied meaning and others. However, it is not clearly illustrating whether there are or how many such sentences with "supra-word meaning" in the neuroimaging stimuli in experiments? In

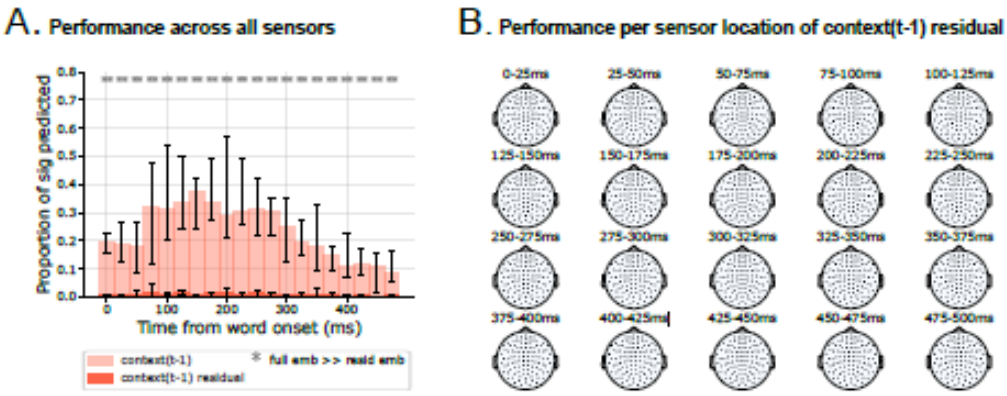

Fig. R4. MEG prediction results at different spatial granularity for embeddings obtained using GPT-2. All subplots present the median across participants and errorbars signify the medians' 95% confidence intervals. **(A)** Proportion of sensors for each timepoint significantly predicted by the full and residual embeddings (visualized in lighter and darker red respectively). Removing the information related to adjacent words from the full context representation results in a significant decrease in performance for all timewindows. No timewindows are significantly different from chance for the residual context embedding. These results replicate the findings using ELMo embeddings that are presented in Fig. 3A. **(B)** Proportions of sensor neighborhoods significantly predicted by each residual embedding. Context-residuals do not predict any sensor-timepoint neighborhood significantly across participants. These results replicate the findings using ELMo embeddings that are presented in Fig. 3B.

the paper, Chapter 9 of Harry Potter novels is taken as neuroimaging stimuli, but it doesn't say what proportions of this type of sentences are contained in Chapter 9? If it's not the majority, then the residual embedding cannot be taken as that encoded the "supra-word meaning".

By our definition, every sentence has some supra-word meaning. Mathematically, we know that the supra-word meaning embedding (i.e. the embedding calculated by context(t-1)-g(word(t), word(t-1),..word(t-n)) contains information because context(t-1) is a non-linear function of word(t-1),..word(t-n), whereas g(word(t), word(t-1),..word(t-n)) is a linear function.

Whether that supra-word meaning is brain-relevant is not known, and this is what we test in the current work. Our results show that the supra-word meaning embedding contains brain-relevant information, because the supra-word meaning embedding predicts significant proportions of several language regions (Fig. 2b in the main paper). We further show in the response to R2 and in Fig. R2 that while the full context embedding from the first hidden layer of ELMo evaluated at word t is strongly linearly related to the next word t+1, current word t, and the previous word t-1, the supra-word embedding for word t does not strongly depend on any individual word, as it was designed to limit contributions of individual words. The presence of the strong linear relationship between the full context representation and the closely adjacent words means that any linear prediction of brain recordings may be overwhelmingly related to these adjacent words. This is also now discussed in the paper in lines 297-312

We have purposefully not made any other assumptions or constraints on the supra-word meaning. We agree that pinpointing exactly what type of supra-word meaning is encoded by ELMo and other NLP models, and how relevant these types are to the brain is an interesting research question that can span many future papers, and we believe that our work is a great stepping stone for this

research direction.

3.There are also some technical details that are not clearly described. Why only the first hidden layer is utilized as the contextual embedding? The second hidden layer also encodes the contextual information. Will the results be different if the second hidden layer or the average of the two layers are utilized?

ELMo has two hidden layers, and we use the first one because of previous work which has shown that the first of two hidden LSTM layers best explains the activity in language regions[16,17]. We also performed experiments with the supra-word meaning obtained from the second hidden layer of ELMo. We did not find any significant differences in the proportion of language regions that are predicted significantly by the supra-word meaning obtained from the first hidden layer vs the supra-word meaning obtained from the second hidden layer. These points are now mentioned in lines 422-426.

Why a ridge regression method is employed to calculate the linear function g? Will there be difference using other method such as one with an analytical solution?

The goal in computing the supra-word meaning is to remove the contribution of individual words to the prediction of brain activity. This is why we use the same type of function to compute the supra-word meaning as we do for predicting the brain activity (i.e. a linear function). Since we use 25 word embeddings and we have 5176 words (12500 input dimensions and 5176 samples), we need some form of regularization to make sure the problem is not under-determined. Using ridge regression to fit a linear model is a simple and standard practice[3,18–20].

4.After obtaining the "supra-word meaning" representations, this paper builds the brain encoding models on two fMRI datasets and one MEG dataset. For the brain encoding method, the following technical details are not clearly explained.

1)To align brain signal with embeddings, this paper chooses to use concatenate vectors of several time steps to model the delay, but the more conventional way is to use HRF. Why authors don't utilize HRF? Are there any special reasons or considerations?

Our approach follows a common solution for building predictive models[9,19,20]. The reason for doing this is that different voxels may exhibit different HRFs so this approach allows for a data-driven estimation of the HRF instead of using the canonical HRF for all voxels. This is now clarified in lines 464-466.

2)For the cross-validation procedure, how to choose the training and validation data? If the training and validation data are too close in time, then their neuroimaging data should be very similar, which will lead to an overfitting result.

The fMRI and MEG Harry Potter data is collected in 4 runs. To estimate all models using this data, we perform 4-fold cross validation. Each fold holds out data corresponding to 1 run that we use to test the generalization of the estimated models. The edges of each run are removed during preprocessing (data corresponding to 20TRs from the beginning and 15TRs from the end of each run), and we further remove data corresponding to additional 5TRs between the training and test data. We follow the same procedures for the movie fMRI data, with the exception that we perform 12-fold cross validation because this dataset was recorded in 12 runs. This is now clarified in lines 438-455.

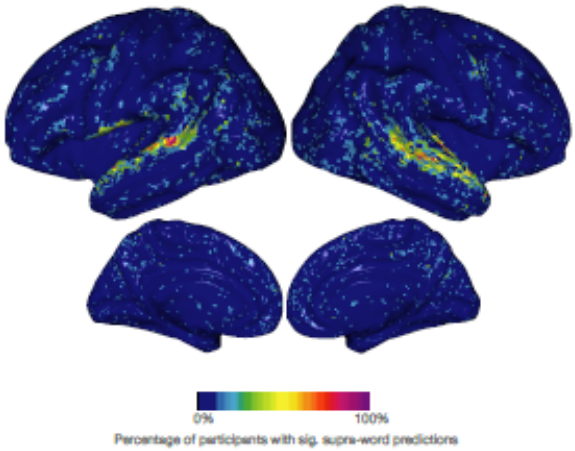

Fig. R5. Group-level significance map of supra-word meaning predictions in the movie fMRI data. We visualize the percentage of all 6 participants in the movie fMRI experiment for which a specific voxel is significantly predicted by ELMo's supra-word meaning (permutation test, followed by FDR correction). We observe consistent supra-word predictive performance in the bilateral ATL and PTL across participants.

3)In the permutation test, why you choose 5TRs as a block?

We perform block permutations so that we can maintain some of the auto-regressive structure of the brain recordings. We used a heuristic to set the block size to 5TRs–because our TR is 2 seconds, we set the block size such that it would include most of a canonical HRF, which peaks around 6 seconds and falls back to baseline over the next several seconds. This is now mentioned in lines 521-524.

4)For the brain encoding results, different subjects show the different patterns especially for the residual context embeddings. Moreover, the anterior and posterior temporal lobes are broad. I am not sure this is enough to verify that "supra-word meaning is predictive of ATL and PTL". How to explain the difference between subjects? Is there a quantitative indicator to calculate a significant group-level results?

We observe very consistent supra-word predictive performance in the bilateral ATL and PTL across participants in the movie fMRI dataset (see Figure R5, now Suppl. Fig. S7). The participants in this study have extensive experience participating in imaging studies (over 100 hours across different paradigms). The differences observed in the Harry Potter dataset may be due to many factors, such as differences in attention levels and fatigue. The participants in the Harry Potter dataset were did not have extensive experience with participating in imaging studies.

5)The results from Harry Potter fMRI and movie fMRI are very different, and there is not any quantitative indicator to explain their similarities and differences. Is it possible to use the method in Figure 4 to calculate the relation between two fMRI dataset? To compare the two results, it is better to give the significant-voxel results (in Figure S1) for the movie data.

We follow the reviewer's suggestion to provide the voxel-level significance maps for the movie

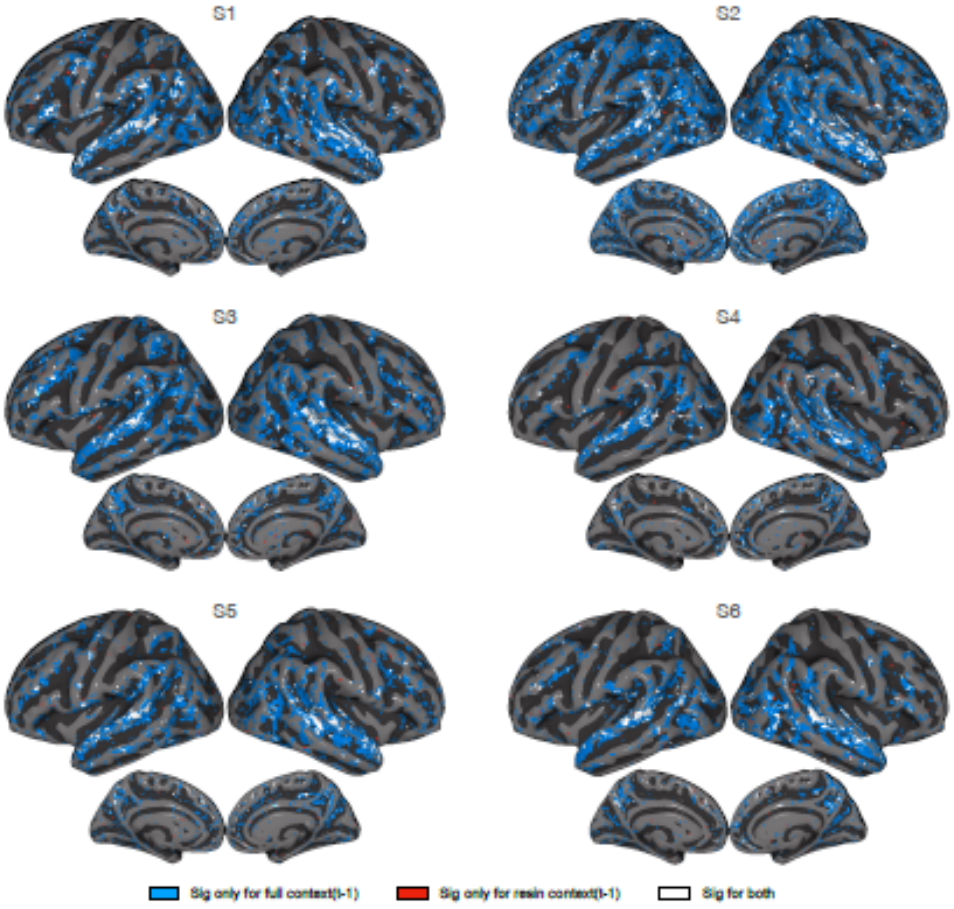

Fig. R6. Individual-level significance maps of prediction performance in the movie fMRI data. Voxels significantly predicted by full-context ELMo embeddings (blue), residual-context ELMo embeddings (red), or both (white), visualized in MNI space. The voxel-level significance is obtained using permutation tests and is FDR corrected. Most of the temporal cortex and IFG is predicted by full context embeddings, with residual context embeddings mostly predicting a subset of those areas.

fMRI dataset, and do so in the Suppl. Fig. S8, and we show them below in Figure R6 for ease of discussion. We would like to clarify that we included the movie fMRI data to provide a replication analysis of supra-word meaning prediction in the language regions. Specifically, the hypothesis derived from the Harry Potter dataset is that the supra-word meaning predicts the ATL and PTL. This result is quantitatively assessed using the bar plot and the significance tests in Fig. 2C and Suppl. Fig. S6. This result is successfully replicated in the movie experiment, where we found that the ATL and the PTL are predicted by supra-word meaning. The movie experiment might have more statistical power because we also found that the supra-word meaning also predicts the IFG orbitalis and posterior cingulate are predicted significantly (with pvalues of 0.045, FDR corrected for multiple comparisons across ROI).

6)The findings of "supra-word meaning is not detectable in MEG" are unexpected and may needs more experiments to verify. There are several factors that could affect the results. For instance, the serial visual format is not a natural way to understand language, and the results may be different when collecting MEG data using a natural audio listening paradigm.

Both the fMRI and MEG data were collected using the same serial visual format, and the supra-word meaning is predictive of a significant proportion of the fMRI recordings, so it is unlikely that the paradigm is the reason for the MEG null result. Nonetheless, our pilot data suggests that supra-word meaning might not predict MEG data during listening either (see the clarification on shared concerns section and the results in Figure R1).

In addition, there are only 8 subjects for the MEG experiments and the individual results are not shown in the paper. Is there any subject having significant results for the supra-word embedding?

We have now included the supra-word meaning results for individual subjects in Suppl. Fig. S13, and include it in Figure R7 below for ease of visualization. This figure is also now referenced in line 160. Of the 6120 total MEG sensor-timepoints (306 sensors x 20 timepoints) per subject, supra-word meaning predicts the following number significantly (at 0.05 level, permutation test followed by FDR correction): 0, 0, 1, 2, 5, 5, 20, 21. Note that these sensor-timepoints do not co-occur across participants. In contrast, the context predicts the following number of sensor-timepoints significantly: 3050, 4203, 3963, 1636, 3010, 5100, 4714, 4702. Given these results, we are confident that removing the individual word information from the context eliminates much of the information that is useful to predict the MEG recordings.

When preprocessing the MEG data, does the time delay between the stimulus and the recording of the actual data be considered?

The time delay between the stimulus and the MEG recording (which was measured during data acquisition) is accounted for during the preprocessing stages. We additionally conducted the supra-word prediction at all stages of preprocessing and did not observe any effect on the (lack of) significance of the results.

Fig. R7. Individual-level proportions of MEG sensor neighborhoods significantly predicted by supra-word meaning. Only the significant proportions are displayed (FDR corrected, $p < 0.05$). Of the 6120 total MEG sensor-timepoints (306 sensors x 20 timepoints) per subject, ELMo's supra-word meaning predicts the following number significantly (at 0.05 level, permutation test followed by FDR correction): 0, 0, 1, 2, 5, 5, 20, 21. In contrast, the context predicts the following number of sensor-timepoints significantly: 3050, 4203, 3963, 1636, 3010, 5100, 4714, 4702. Removing the individual word information from the context eliminates much of the information that is useful to predict the MEG recordings.

References

1. Chen, G. *et al.* Hyperbolic trade-off: the importance of balancing trial and subject sample sizes in neuroimaging. *NeuroImage* **247,** 118786 (2022).
2. Schoffelen, J.-M. *et al.* A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific data* **6,** 1–13 (2019).
3. Huth, A. G. *et al.* Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532,** 453–458 (2016).
4. Herholz, P. *et al.* A roadmap to reverse engineering real-world generalization by combining naturalistic paradigms, deep sampling, and predictive computational models. *arXiv preprint arXiv:2108.10231* (2021).
5. Goldman-Rakic, P. S. Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences* **93,** 13473–13480 (1996).
6. Luck, S. J., Vogel, E. K. & Shapiro, K. L. Word meanings can be accessed but not reported during the attentional blink. *Nature* **383,** 616–618 (1996).
7. Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386,** 608–611 (1997).
8. Goldstein, A. *et al.* Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv,* 2020–12 (2021).
9. Schrimpf, M. *et al.* Artificial Neural Networks Accurately Predict Language Processing in the Brain. *BioRxiv* (2020).
10. Caucheteux, C. & King, J.-R. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv* (2020).
11. Toneva, M., Williams, J., Bollu, A., Dann, C. & Wehbe, L. *Same Cause; Different Effects in the Brain* in *First Conference on Causal Learning and Reasoning* (2021).
12. Partee, B. Noun phrase interpretation and type-shifting principles. *Formal semantics: The essential readings,* 357–381 (2002).
13. Pylkkänen, L. Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B* **375,** 20190299 (2020).
14. Chelba, C. *et al.* One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling (2014).
15. Peters, M. E. *et al. Deep contextualized word representations* in *Proceedings of NAACL-HLT* (2018), 2227–2237.
16. Jain, S. & Huth, A. *Incorporating context into language encoding models for fmri* in *Advances in neural information processing systems* (2018), 6628–6637.
17. Toneva, M. & Wehbe, L. *Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)* in *Advances in Neural Information Processing Systems* (2019), 14928–14938.

18. Mitchell, T. *et al.* Predicting human brain activity associated with the meanings of nouns. *Science* **320,** 1191–1195 (2008).

19. Nishimoto, S. *et al.* Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* (2011).

20. Wehbe, L. *et al.* Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses. *PloS one* **9** (ed Paterson, K.) e112575. ISSN: 1932-6203. http://dx.plos.org/10.1371/journal.pone.0112575 (Nov. 2014).

Decision Letter, first revision:

Date: 23rd May 22 17:10:55
Last Sent: 23rd May 22 17:10:55
Triggered By: Kaitlin McCardle
From: kaitlin.mccardle@us.nature.com
To: lwehbe@cmu.edu
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-21-0986A
Message: ** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Dr Wehbe,

Your manuscript "Combining computational controls with natural text reveals new aspects of meaning composition" has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address all of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Please be sure to add additional clarifications of the supra-word meanings, supra-word embeddings and what they both capture, as you discussed in your previous email.
- As you have noted in your previous email, fully understanding the meaning behind your measures may require a large experimental undertaking, and may be out of the scope of your current goals. Please be sure to clarify this in your revision (and also note future research in this direction in your discussion section)
- Please update your code repository to reflect any new experiments.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making of use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best regards,

Kaitlin McCardle
Editor
Nature Computational Science

Reviewers comments:

Reviewer #1 (Remarks to the Author):

I would like to thank the authors for considering my comments and for their responses to the comments. I agree with the authors hat their manuscript has benefited from the revisions. Nonetheless, I still have some remaining suggestions as outlined below.

I initially said that I found it irritating that the full context embedding was predictive of much of the fMRI recordings across the brain. The authors in response changed their wording to say that the full context embedding was predictive of "fMRI recordings in the language and semantic networks". However, I would find it important to see some demonstration of the specificity of the prediction performance to the brain's language network.

Relatedly, in Fig. S15 it looks like that there are significant effects of the full context already in the 0-25ms and 25-50ms time segments, which seems very implausible. Is this correct and if so, what does it mean?

Finally, I agree with both of my fellow reviewers' request that it would be useful and important that the authors demonstrate that their computational measure of supra-word meaning captures relevant aspects of meaning. The authors say in their response that this should be left for future work, but I think (also given that there is quite some variability in what this measure predicts across participants and datasets, and e.g., the "group level significance map" displayed in Fig. R5 does not actually show group level statistics, but rather percentage of participant level significance for 6 participants) that it would be important to show that the measure captures interesting aspects of meaning. (Concerning my earlier comment that it would be important to replicate the results in a larger MEG data set I still think so but I accept the authors' response that there is no such data set openly available and I think that given that their expertise might be more in the computational field than in MEG data acquisition it would be too much to ask them to acquire these MEG data themselves. However, some demonstration of what their measure actually captures seems well in line with the authors' expertise so I think this is a reasonable request by my fellow reviewers and not too much to ask, especially for a Nature journal.)

Reviewer #2 (Remarks to the Author):

The aim of this paper is to study sentence meaning at the brain level. The aim is to go beyond the "bag of words" that make up the sentence and to get at processing beyond that. The authors take a data-driven perspective, using recent word and sentence embeddings from NLP to match to either fMRI or MEG signals. In order to abstract from the "bag of words" (and remove the very strong signal from the individual words), they define a new computational representation that is basically the residual of embeddings from NLP with respect to the immediate context words. One of the most intriguing results is a negative result: that these new supra-word embeddings can predict fMIR but not MEG signals that they obtained.

In their revision of the paper, the authors have significantly strengthened their

negative result:
They have added another computational experiment using a second model of contextualized embeddings, GPT-2 in addition to their previous ELMO, showing that the result stays the same.
They have also added another MEG dataset, from a single subject listening to the Radio Hour, showing that their result still holds there.
They also argue that there is no problem with their new supra-word embeddings since it is possible to predict fRMI signals from them, and there is no problem with their MEG signals since it is possible to predict them with whole-context embeddings.

This all makes sense, and in my opinion is sufficient to show that their negative result is real and interesting.

The authors have also added an analysis of their supra-word embeddings to shed light on what these embeddings are -- this is great, and is what I had asked them to do. They show the importance of individual words for predicting either the original (whole-context) embeddings or residual (supra-word) embeddings. The immediate context is of overwhelming importance for a linear model of whole-context embeddings. For the residual embeddings, no single word is of high importance.

However, one main point of confusion, which had been noted by both me and reviewer 3, remains: It remains unclear what the authors mean by "supra-word meaning", and this confusion makes it unclear what their overall argument is.

In their rebuttal, and in the abstract, they define 'supra-word meaning' to be a technical object, a computational model: "we adopt a data-driven computational approach to study the combined meaning of words beyond their individual meaning. We term this product “supra-word meaning”". So here, supra-word meaning is the same as residual embedding. It is not a linguistic object, it is a computational object, a type of embedding. But in the introduction, they use the term "supra-word meaning" to designate a *linguistic* object: "For
13 example, we understand the statement that “Mary finished the apple” to mean that Mary finished eating
14 the apple, even though “eating” is not explicitly specified1. This supra-word meaning, or the product of
15 meaning composition beyond the meaning of individual words, is at the core of language comprehension,
16 and its neurobiological bases and processing mechanisms must be specified in the pursuit of a complete
17 theory of language processing in the brain." Reviewer-3 takes them at their word and asks whether *linguistically* they can show that their data contains what they call "supra-word meaning" here, for example implied meaning. in the rebuttal, the authors state that by their definition, every sentence has supra-word meaning -- but this is the case only by definition has supra-word meaning in sense 1, a residual embedding. Reviewer-3 meant sense 2 of "supra-word meaning" -- but they are right, not every sentence necessarily contains implicatures, or cases of regular polysemy such as "Mary finished the apple".

So what the authors show is *not* that every sentence has supra-word meaning as a linguistic object, or that there is linguistic supra-word meaning processing going on in the brain for these specific phenomena listed in the introduction (which, in their

current form, are highly misleading). They have computed a new type of embedding, a residual embedding, which for now we'll call XYZ. This XYZ manages to predict fMRI patterns in language understanding centers quite well, but not MEG patterns. What this XYZ is, is unclear. From the way it is computationally derived, it has something to do with whole-sentence meaning representations, and it stands to reason that because the original whole-context embeddings are highly effective at many language processing tasks, including semantic tasks, XYZ also contains information about meaning. XYZ also intriguingly acts differently when used to predict fMRI and MEG, which means that we should study XYZ further. It is possible that XYZ has something to do with concept combination, because of the way it was computed as a residual, but that is speculative right now and needs further analysis.

Anyway, what I mean by this is: The main argument of the paper still needs clarification, and careful separation of the two senses of supra-word meaning that the authors have been using.

With regard to the previously raised point:
2. A related question is does the embedding calculated by context(t-1)-g(word(t), word(t-1)) really capture the “supra-word meaning” information? This hypothesis has never been testified in the paper. This is very important and should be testified before the experiments of brain encoding. If it doesn’t, then other experiment will be the buildings in air. It is meaningless. There are some ways to testify this such as to build a test dataset including phrases with “supra-word meaning” along with some human annotations. Moreover, in the introduction, the “supra-word meaning” is defined as the composed meaning of a sequence of words that is not part of the corresponding bag-of-words, i.e., implied meaning and others. However, it is not clearly illustrating whether there are or how many such sentences with “supra-word meaning” in the neuroimaging stimuli in experiments? In the paper, Chapter 9 of Harry Potter novels is taken as neuroimaging stimuli, but it doesn’t
say what proportions of this type of sentences are contained in Chapter 9? If it's not the majority, then the residual embedding cannot be taken as that encoded the “supra-word meaning.

This, I think, is the issue that I focused on in my review: They define "supra-word meaning" on the one hand as a mathematical object, obtained as residual of word embeddings with respect to the task of predicting word context. Call this supra-word meaning 1. And on the other hand they refer to "supra-word meaning" as something linguistic/mental, something that people compute, which includes implicit meaning and which involves concept combination effects. Call this supra-word meaning 2. But they never actually show that supra-word meaning 1 models supra-word meaning 2. This is what has confused reviewer 3 here.

I think the argument that the authors try to make is: supra-word meaning 2 is too hard to define and work with, so they focus on supra-word meaning 1, which is straightforward to compute. It might have something to do with supra-word meaning 2 because of the way it is computed, but that is unclear and is beyond the paper's scope to show. But the one thing that really makes supra-word meaning 1 interesting is the experimental finding in predicting fMRI vs MEG, so we should analyze supra-word meaning 1 further.

This is not an exceedingly strong argument, but it might be good enough -- if the

authors could make it clearly enough

I just took a quick look at the code repository. It looks like the README is comprehensive, with instructions on required packages and exact calls. Their code is reasonably readable, too. I don't have time to run the code at this time, but I would say this looks good with respect to reproducibility.

Reviewer #4 (Remarks to the Author):

- Let me start by saying that this paper presents, in my opinion, some very interesting measures and results, and I hope to see it published soon in some form.
- Reviewers in the first round already went into much detail regarding the experimental support for the claims. In my assessment, the authors did a good job in addressing these concerns. The GPT2 results and the additional controls really strengthen the message, and the additional discussion of caveats in the Discussion section will help readers judge how strong the results really are.
- The one remaining problem, in my view, with the paper is about presentation. I found the order in which information is offered not always ideal, and the introduction and the used terminology risk antagonizing some readers unnecessarily. If the authors manage to improve the paper's framing, it will have a better chance of reaching the broad readership it deserves.
- It's difficult to give very concrete suggestions for improvements, but here are a few ideas:
* The key concept in the paper now is called 'supra-word meaning'. That's an awkward term, and the authors further complicate the matter by proposing it as a formalization of a linguistic intuition, defining/describing it multiple times rather imprecisely and leaving a more technical definitions to a figure. Wouldn't it be easier to simply first present it as a technical concept with a technical name (e.g. something with the terms "word pair"/ linear / residual in it), and show that it yields a very interesting pattern of results in predicting brain data. The challenge of interpreting the concept in linguistic terms can then be presented later, with the caveat that it is far from solved.
* Regardless of what it is called, define it properly the first time in the main text. The "definition" on line 42 is difficult to understand and different from the graphical definition in fig 1 (lack of linear predictability is not the same as bringing in new aspects of meaning: the meanings of "two hundred" and "kick the bucket" are both not linearly predictable, but for very different reasons).
* The introduction now reads somewhat dismissive of linguistic theory, while citing linguistic empirical work (Pylkkanen et al) and a rather random selection of old and new theoretical papers. By simply stating that linguistic theory generally has had very different goals than what is required for this paper, some unnecessary controversy could be avoided. Also, perhaps Baroni & Linzen's "On the proper role of linguistically-oriented deep net analysis in linguistic theorizing", and the way they refer to linguistic theory, is useful here.
* The most interesting finding in the paper is the null results for MEG, in combination with its controls. It's odd that the introduction doesn't talk about the reasons why that is so interesting (the long standing questions about relating MRI to MEG).

Finally, maybe it's just me (and the time pressure), but I didn't understand how the method for computing spatial generalization matrices deals with the fact that different

brain areas will use very different vector spaces even if they are encoding similar information (unlike the situation in temporal generalization matrices).

Author Rebuttal, first revision:

Author Response: Combining computational controls with natural text reveals new aspects of meaning composition

We thank the reviewers and editors for their time in reviewing our response and the edited manuscript, and for their additional feedback. We were encouraged by the reviewers' positive responses to our efforts in addressing the previous round of feedback. We believe that incorporating the reviewers' new suggestions has further strengthened the manuscript. We further respond to several important comments that are shared between reviewers in Section 1. We lastly include point-by-point responses to the remaining comments of the reviewers in Section 2 with references to changes in the manuscript. Line numbers cited here correspond to lines in the version of the manuscript labeled with changes.

1 Response to shared concerns

Several reviewers point out that the introduction of supra-word meaning is confusing and its motivation needs to be clarified to strengthen our argument. We quote the comments below:

Reviewer 2: "However, one main point of confusion, which had been noted by both me and reviewer 3, remains: It remains unclear what the authors mean by "supra-word meaning", and this confusion makes it unclear what their overall argument is.

In their rebuttal, and in the abstract, they define 'supra-word meaning' to be a technical object, a computational model: "we adopt a data-driven computational approach to study the combined meaning of words beyond their individual meaning. We term this product "supra-word meaning"". So here, supra-word meaning is the same as residual embedding. It is not a linguistic object, it is a computational object, a type of embedding. But in the introduction, they use the term "supra-word meaning" to designate a *linguistic* object: "For 13 example, we understand the statement that "Mary finished the apple" to mean that Mary finished eating 14 the apple, even though "eating" is not explicitly specified1. This supra-word meaning, or the product of 15 meaning composition beyond the meaning of individual words, is at the core of language comprehension, 16 and its neurobiological bases and processing mechanisms must be specified in the pursuit of a complete 17 theory of language processing in the brain." Reviewer-3 takes them at their word and asks whether *linguistically* they can show that their data contains what they call "supra-word meaning" here, for example implied meaning. in the rebuttal, the authors state that by their definition, every sentence has supra-word meaning – but this is the case only by definition has supra-word meaning in sense 1, a residual embedding. Reviewer-3 meant sense 2 of "supra-word meaning" – but they are right, not every sentence necessarily contains implicatures, or cases of regular polysemy such as "Mary finished the apple".

[...]

Anyway, what I mean by this is: The main argument of the paper still needs clarification, and careful separation of the two senses of supra-word meaning that the authors have been using.

[...]

This, I think, is the issue that I focused on in my review: They define "supra-word meaning" on the one hand as a mathematical object, obtained as residual of word embeddings with respect to the task of predicting word context. Call this supra-word meaning 1. And on the other hand they refer to "supra-word meaning" as something linguistic/mental, something that people compute, which includes implicit meaning and which involves concept combination effects. Call this supra-word meaning 2. But they never actually show that supra-word meaning 1 models supra-word meaning 2. This is what has confused reviewer 3 here.

I think the argument that the authors try to make is: supra-word meaning 2 is too hard to define and work with, so they focus on supra-word meaning 1, which is straightforward to compute. It might have something to do with supra-word meaning 2 because of the way it is computed, but that is unclear and is beyond the paper's scope to show. But the one thing that really makes supra-word meaning 1 interesting is the experimental finding in predicting fMRI vs MEG, so we should analyze supra-word meaning 1 further.

This is not an exceedingly strong argument, but it might be good enough – if the authors could make it clearly enough."

Reviewer 4: "The key concept in the paper now is called 'supra-word meaning'. That's an awkward term, and the authors further complicate the matter by proposing it as a formalization of a linguistic intuition, defining/describing it multiple times rather imprecisely and leaving a more technical definitions to a figure. Wouldn't it be easier to simply first present it as a technical concept with a technical name (e.g. something with the terms "word pair"/ linear / residual in it), and show that it yields a very interesting pattern of results in predicting brain data. The challenge of interpreting the concept in linguistic terms can then be presented later, with the caveat that it is far from solved."

The reviewers' comments are very helpful for understanding the confusion. We have updated the introduction to more clearly differentiate the concepts of "supra-word meaning" and "supra-word embedding" (i.e. "residual embedding" which is how we refer to it in the Figures) in lines 37-76, and have updated the rest of the manuscript to maintain this distinction.

- *The "supra-word meaning" of a sentence is **latent** and may relate to **a set** of linguistic or psychological objects. In the current introduction, we speculate about different types of sentences or relationships between words that may result in supra-word meaning. We now clarify in lines 37-71 that we don't intend to study one specific linguistic object, but rather to enable a data-driven study of sentence effects that are beyond the word-level.*

- *The "supra-word embedding" or "residual embedding" that we construct is a computational object. We can use this technical object to shed light on the different processes that underlie the latent supra-word meaning. The problem can be defined technically as constructing an embedding for the variance encoded in the full context embedding that is orthogonal to the individual word embeddings. Then, we can use this "supra-word embedding" computational object to evaluate where and when in the brain does the variance captured in the full context embedding but not the individual words predict activity. We clarify this in lines 71-76.*

- *This "supra-word embedding" is an interesting construct for 2 reasons: 1) it contains brain-relevant information because it explains significant proportions of fMRI recordings in 2 specific language regions that have been previously related to language composition, and 2) as already pointed out by R2, the supra-word embedding predicts fMRI but not MEG which is an interesting experimental finding only enabled by the construction of the supra-word embedding.*

Another shared question is whether the constructed supra-word embedding contains aspects of brain-relevant meaning, and if so, what are these aspects.

Reviewer 1: "I agree with both of my fellow reviewers' request that it would be useful and important that the authors demonstrate that their computational measure of supra-word meaning captures relevant aspects of meaning. The authors say in their response that this should be left for future work, but I think (also given that there is quite some variability in what this measure predicts across participants and datasets, and e.g., the "group level significance map" displayed in Fig. R5 does not actually show group level statistics, but rather percentage of participant level significance for 6 participants) that it would be important to show that the measure captures interesting aspects of meaning. (Concerning my earlier comment that it would be important to replicate the results in a larger MEG data set I still think so but I accept the authors' response that there is no such data set openly available and I think that given that their expertise might be more in the computational field than in MEG data acquisition it would be too much to ask them to acquire these MEG data themselves. However, some demonstration of what their measure actually captures seems well in line with the authors' expertise so I think this is a reasonable request by my fellow reviewers and not too much to ask, especially for a Nature journal.)"

Reviewer 2: "So what the authors show is *not* that every sentence has supra-word meaning as a linguistic object, or that there is linguistic supra-word meaning processing going on in the brain for these specific phenomena listed in the introduction (which, in their current form, are highly misleading). They have computed a new type of embedding, a residual embedding, which for now we'll call XYZ. This XYZ manages to predict fMRI patterns in language understanding centers quite well, but not MEG patterns. What this XYZ is, is unclear. From the way it is computationally derived, it has something to do with whole-sentence meaning representations, and it stands to reason that because the original whole-context embeddings are highly effective at many language processing tasks, including semantic tasks, XYZ also contains information about meaning. XYZ also intriguingly acts differently when used to predict fMRI and MEG, which means that we should study XYZ further. It is possible that XYZ has something to do with concept combination, because of the way it was computed as a residual, but that is speculative right now and needs further analysis.

[...]

With regard to the previously raised point: 2. A related question is does the embedding calculated by context(t-1)-g(word(t), word(t-1)) really capture the "supra-word meaning" information? This hypothesis has never been testified in the paper. This is very important and should be testified before the experiments of brain encoding. If it doesn't, then other experiment will be the buildings in air. It is meaningless. There are some ways to testify this such as to build a test dataset including phrases with "supra-word meaning" along with some human annotations. Moreover, in the introduction, the "supra-word meaning" is defined as the composed meaning of a sequence of words that is not part of the corresponding bag-of-words, i.e., implied meaning and others. However, it is not

clearly illustrating whether there are or how many such sentences with "supra-word meaning" in the neuroimaging stimuli in experiments? In the paper, Chapter 9 of Harry Potter novels is taken as neuroimaging stimuli, but it doesn't say what proportions of this type of sentences are contained in Chapter 9? If it's not the majority, then the residual embedding cannot be taken as that encoded the "supra-word meaning."

*Summary: We understand and share the desire to exactly pinpoint what linguistic or psychological information relates to the significant prediction of the bilateral ATL and PTL that we observe using the supra-word embeddings. However, even if we understand that information X,Y, and Z is contained in the supra-word embeddings, that **will not** answer this question. In contrast, we can use methods that we propose in the current work to answer this question, but that will require collecting several new dataset annotations for the Harry Potter stimulus text and a large additional experimental undertaking. We have updated the Discussion in lines 381-404 to more clearly reflect this. We give more details below:*

- *Firstly, we believe that the fact that the supra-word embedding is able to predict significant proportions of two language understanding regions is in itself a validation of its ability to capture brain-relevant aspects of meaning. We also want to point out that there are several other works that have been published recently in general audience venues that relate brain recordings to neural network language model embeddings without knowing exactly what linguistic or psychological information is captured by these embeddings, such as Schrimpf et al. (2021) in PNAS[1], and Goldstein et al. (2022) in Nature Neuroscience[2].*

- *That said, we understand and share the desire to exactly pinpoint what linguistic or psychological information relates to the significant prediction of the bilateral ATL and PTL that we observe using the supra-word embeddings. However, we want to emphasize that even if we understand that information X,Y, and Z is contained in the supra-word embeddings, that **will not** answer this question.*

- *The reason is that the supra-word embedding contains X, Y, Z, and possibly other information W, that we have not measured, and any one of them may be the information that is predictive of the bilateral ATL and PTL. For example, as our work shows, the MEG recordings are predicted by the full context embedding, but not by the information in the full context embedding that is orthogonal to the individual word embeddings (i.e. supra-word embeddings). Therefore, only knowing 2 things—1) full context embeddings predict MEG, and 2) some supra-word information is contained in the full context embeddings—is not enough to reveal what information in the full context embeddings is predictive of the MEG recordings, and suggesting that it may be because of the supra-word information would be misleading. Similarly, linguistic information X,Y, and Z may be contained in the supra-word embedding, but may not be necessary for predicting the bilateral ATL and PTL.*

- *This is in fact one of the central points of our work, and we believe it is critical to make this point as it is becoming increasingly popular to relate brain recordings to complex neural network-derived embeddings that contain multiple sources of information.*

- *In this work, we show one way forward from this by utilizing computational controls. If linguistic information X,Y, and Z are shown to be contained in the supra-word embedding,*

one could regress all combinations of the corresponding linguistic labels from the supra-word embedding and observe how the prediction of fMRI recordings in the bilateral ATL and PTL changes as a result.

- *These analyses are not simple and they would require the Harry Potter stimulus dataset (5100 words) to be annotated with various linguistic and psychological labels, some of which may require expert linguists (e.g. dependencies between words that follow from different semantic formalisms) who might even disagree amongst themselves. It's an important undertaking and certainly a next step in this research direction, but we estimate that it will take 6 months of effort for a lab focused on NLP to collect this additional annotation data and conduct all the required analyses.*

- *We hope that we have conveyed the care with which we think about these issues of interpretability and our commitment to creating methods that can benefit from recent progress in large neural network models while also granting us more control over the scientific inferences we can make.*

2 Point-by-point responses to reviewers

2.1 Reviewer 1

I would like to thank the authors for considering my comments and for their responses to the comments. I agree with the authors hat their manuscript has benefited from the revisions. Nonetheless, I still have some remaining suggestions as outlined below.

We thank Reviewer 1 for their comment.

I initially said that I found it irritating that the full context embedding was predictive of much of the fMRI recordings across the brain. The authors in response changed their wording to say that the full context embedding was predictive of "fMRI recordings in the language and semantic networks". However, I would find it important to see some demonstration of the specificity of the prediction performance to the brain's language network.

In order to demonstrate the specificity of the context vector to the language network, we drew the Fedorenko ROIs on the MNI surface, and plotted the average prediction performance across the subjects in figure 2 for the full context vector. We find that our results (shown here in response Fig. R1) align with our expectations from the literature (e.g., Huth et al 2016, Caucheteux et al 2021, shown in response Fig. R2). Namely, the regions that are well predicted span the language network along with some additional regions in the frontal cortex and lateral parietal cortex.

Relatedly, in Fig. S15 it looks like that there are significant effects of the full context already in the 0-25ms and 25-50ms time segments, which seems very implausible. Is this correct and if so, what does it mean?

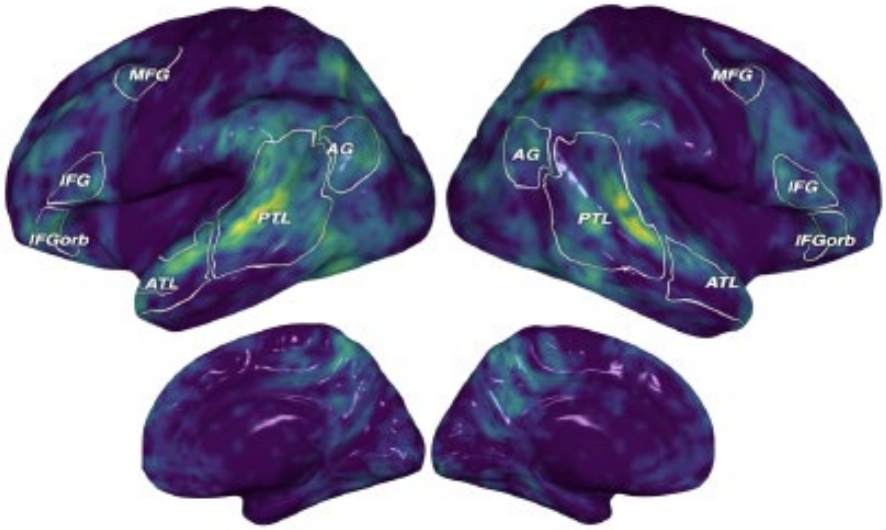

Fig. R1. Average prediction performance across the four subjects in Fig. 2 for the context vector. This result is also part of supplementary Fig. S2.

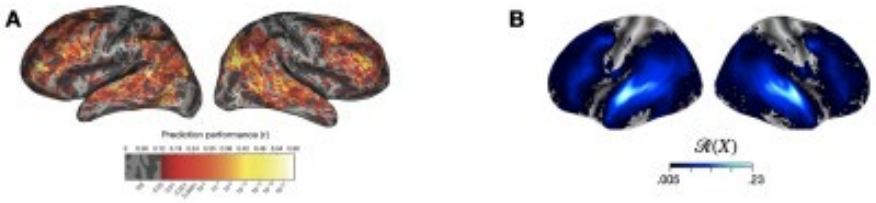

Fig. R2. Examples from the literature of regions well predicted by semantic properties of narratives. [A] was adapted from Huth et al. (2016) [3] and [B] from Caucheteux et al. (2021) [4]. These figures are only used for the purpose of this response.

We attribute the significant predictive performance of the full context in all time windows in MEG to information related to the individual words, since there is no remaining significant performance once we remove the individual word information from the full context embeddings. Specifically, the significant predictive performance in the early time windows during the presentation of word t may be related to the prediction of word t+1, and/or to continual processing of the previous word t-1, since information about both word t-1 and word t is contained in the full context embedding. The results in Fig. 3A suggest that both types of information, independently, are predictive of these early time windows. We now clarify this in the caption of Supplementary Figure S16 (prior Figure S15).

2.2 Reviewer 2

The aim of this paper is to study sentence meaning at the brain level. The aim is to go beyond the "bag of words" that make up the sentence and to get at processing beyond that. The authors take a data-driven perspective, using recent word and sentence embeddings from NLP to match to either fMRI or MEG signals. In order to abstract from the "bag of words" (and remove the very strong signal from the individual words), they define a new computational representation that is basically the residual of embeddings from NLP with respect to the immediate context words. One of the most intriguing results is a negative result: that these new supra-word embeddings can predict fMIR but not MEG signals that they obtained.

In their revision of the paper, the authors have significantly strengthened their negative result: They have added another computational experiment using a second model of contextualized embeddings, GPT-2 in addition to their previous ELMO, showing that the result stays the same. They have also added another MEG dataset, from a single subject listening to the Radio Hour, showing that their result still holds there. They also argue that there is no problem with their new supra-word embeddings since it is possible to predict fRMI signals from them, and there is no problem with their MEG signals since it is possible to predict them with whole-context embeddings.

This all makes sense, and in my opinion is sufficient to show that their negative result is real and interesting.

The authors have also added an analysis of their supra-word embeddings to shed light on what these embeddings are – this is great, and is what I had asked them to do. They show the importance of individual words for predicting either the original (whole-context) embeddings or residual (supra-word) embeddings. The immediate context is of overwhelming importance for a linear model of whole-context embeddings. For the residual embeddings, no single word is of high importance.

We thank reviewer 2 for these comments.

The rest of reviewer 2's comments are included in the shared section above.

We thank Reviewer 2 for the very detailed and constructive feedback regarding the two concerns discussed in Section 1. We have tried to address these in the revised manuscript.

2.3 Reviewer 4

- Let me start by saying that this paper presents, in my opinion, some very interesting measures and results, and I hope to see it published soon in some form. - Reviewers in the first round already went

into much detail regarding the experimental support for the claims. In my assessment, the authors did a good job in addressing these concerns. The GPT2 results and the additional controls really strengthen the message, and the additional discussion of caveats in the Discussion section will help readers judge how strong the results really are.

We thank reviewer 4 for these comments.

- The one remaining problem, in my view, with the paper is about presentation. I found the order in which information is offered not always ideal, and the introduction and the used terminology risk antagonizing some readers unnecessarily. If the authors manage to improve the paper's framing, it will have a better chance of reaching the broad readership it deserves.

[...]

* Regardless of what it is called, define it properly the first time in the main text. The "definition" on line 42 is difficult to understand and different from the graphical definition in fig 1 (lack of linear predictability is not the same as bringing in new aspects of meaning: the meanings of "two hundred" and "kick the bucket" are both not linearly predictable, but for very different reasons).

We thank the reviewer for helping us improve the presentation of the work. We now define the technical concept "supra-word embedding" more precisely in 71-76. We further hope that separating the concepts of "supra-word meaning" and "supra-word embedding" has also helped with this concern.

The introduction now reads somewhat dismissive of linguistic theory, while citing linguistic empirical work (Pylkkanen et al) and a rather random selection of old and new theoretical papers. By simply stating that linguistic theory generally has had very different goals than what is required for this paper, some unnecessary controversy could be avoided. Also, perhaps Baroni & Linzen's "On the proper role of linguistically-oriented deep net analysis in linguistic theorizing", and the way they refer to linguistic theory, is useful here.

We appreciate the feedback and the reviewer's suggestion. We replaced the previous discussion of linguistic theory with a new discussion in lines 77-85, which follows Baroni & Linzen's example. We hope that by clarifying the notions of supra-word meaning and supra-word embedding, and how our goals differ from those of traditional linguistic theory, we have resolved this problem.

The most interesting finding in the paper is the null results for MEG, in combination with its controls. It's odd that the introduction doesn't talk about the reasons why that is so interesting (the long standing questions about relating MRI to MEG).

We are very happy that the reviewer is supportive of the importance of the null findings in MEG. According to the reviewer's suggestion, we now include a summary of the debate about relating fMRI to MEG in the introduction in lines 110-119.

Finally, maybe it's just me (and the time pressure), but I didn't understand how the method for computing spatial generalization matrices deals with the fact that different brain areas will use very different vector spaces even if they are encoding similar information (unlike the situation in tempo-

ral generalization matrices).

Perhaps what the reviewer is alluding to is that two regions might be sensitive to the meaning of words, but represent aspects of this meaning in a different basis. To give an illustration of this example in the vision domain, we can imagine two regions that are sensitive to faces, but one encodes the low level edge and curvature properties of faces and the other one encodes high level properties of faces such as expression. What would the spatial generalization metric measure between these two regions?

In our setting, we can assume there exists a global, large semantic space that encodes semantic information, with the understanding that which semantic dimensions are encoded by each region, and the way in which they are encoded, can differ. We use as an approximation of the semantic space the supra-word embedding, and the way that a brain region encodes the dimensions in that space corresponds to the estimated weights of the encoding model. It is most likely that a brain region will be encoding information using a different semantic basis than our neural network derived embedding, and two brain regions might not be using the same basis. Note that the spatial generalization takes into consideration only regions that are already significantly predicted by the supra-word embedding. Since the two hypothetical regions are predicted by the encoding model using the supra-word embedding, we can assume that the supra-word embedding captures some of the dimensions spanned by their bases, up to a transformation. The difference between the bases of the two regions could manifest as some difference in the weights of the learned encoding models for the two regions. Therefore spatial generalization allows us to test how similar the semantic sensitivity of two regions are by measuring how well the estimated encoding of one voxel transfers to another voxel. Note that this is similar to methods that directly compare the encoding model weights estimated for two voxels[35,36], but spatial generations looks not only at how much the weights are different, but also at how much this difference affects the ability to predict held-out data. It could be considered a more stringent method, ignoring small differences in weights that don't affect generalization performance. We have included this clarification in Methods and Materials in lines 600-618.

References

1. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118** (2021).
2. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nature neuroscience* **25,** 369–380 (2022).
3. Huth, A. G., Heer, W. A. D., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532,** 453–458 (2016).
4. Caucheteux, C., Gramfort, A. & King, J.-R. Long-range and hierarchical language predictions in brains and algorithms. *arXiv preprint arXiv:2111.14232* (2021).
5. Çukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience* **16,** 763–770 (2013).
6. Deniz, F., Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience* **39,** 7722–7736 (2019).

nature research

Decision Letter, second revision:

Date: 1st September 22 14:58:46
Last Sent: 1st September 22 14:58:46
Triggered By: Kaitlin McCardle
From: kaitlin.mccardle@us.nature.com
To: lwehbe@cmu.edu
CC: computationalscience@nature.com
Subject: AIP Decision on Manuscript NATCOMPUTSCI-21-0986B
Message: Our ref: NATCOMPUTSCI-21-0986B

1st September 2022

Dear Dr. Wehbe,

Thank you for submitting your revised manuscript "Combining computational controls with natural text reveals new aspects of meaning composition" (NATCOMPUTSCI-21-0986B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science Please do not hesitate to contact me if you have any questions.

Sincerely,

Kaitlin McCardle
Editor
Nature Computational Science

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #2 (Remarks to the Author):

The revisions that the authors have made have, to me, greatly clarified the paper and its main points, and I am now happy with the manuscript as it is.

The clear distinction that the authors now make throughout between supra-word meaning and supra-word embeddings is extremely helpful and makes the overall argument much easier to follow.

I particularly like the new paragraph from lines 49 through 79, which provides a strong motivation for the use of neural language models in studying supra-word meaning, and an insightful statement on what it is that we can or cannot learn from the use of language models. Great point about the usefulness of a single, holistic supra-word embedding as opposed to testing an individual linguistic phenomenon.

I also really like the addition of lines 395-400, where the authors point out why it would not be enough to know that linguistic phenomena X, Y, Z are included in supra-word embeddings, and how the residuals method can be used in the future. I had not understood before that the residual method can in principle be used to *remove* traces of particular linguistic phenomena from the supra-word embedding for more in-depth tests.

This is an important point, I think, and an important contribution to the debate on how we should be using language models in the future.

As I think I already said in my previous rounds of comments, it is good that the authors re-ran their analyses on additional datasets (the movie data, and the radio stories data) and with an additional language model with a different architecture (GPT-2), to show that the null result on MEG data is a true finding, and not a problem with the data.

The code uses Python, a widely used and easily accessible programming language.

There are few Python packages required to run the code, and they are all, again, widely used and easily accessible. There was only one additional package I needed to install, and that was installed in less than a minute.

The README is clear and well-organized, and gives the exact commands needed to run, along with expected output. The code itself comes in three short and well organized Python scripts. It has few comments, but the function names are well chosen so one can guess what each function does.

I ran the first three commands suggested in the README. The very first run gave me an error, FileNotFoundError: [Errno 2] No such file or directory: './results/predict_fmri_I_gpt2_with_prev-1-context.npy'. But when I made an empty subdirectory called "results" from the directory with the code, and then re-ran, everything ran without a hitch, and produced exactly the results given in the README.

Concerning the third paragraph of Reviewer 1's comments, I think the authors have addressed this point, which coincides with comments that I made and that also other reviewers made. They have carefully separated the terms "supra-word meaning" and "supra-word embedding", and they have added, both in their rebuttal and in the main text, a paragraph that I find quite important. It suggests using embeddings to compare to brain patterns exactly because they seem to be based on a mixture of all sorts of different linguistic phenomena

-- Katrin Erk

Reviewer #4 (Remarks to the Author):

I have enjoyed reading the second revision of the paper. As I wrote in my earlier review, I think the paper represents an important contribution to science. There have been a lot of papers in the last few years showing that vectors extracted from deep language models (LMs) can be used to predict brain activation in MRI, EEG and MEG. This paper attempts to make an important advance by disentangling various sources of information.

To this end, the paper defines 'context residuals': the difference between the representation of prior context of a given word that one can directy extract from an LM (before seeing the target word), and the representation of context that one would retroactively expect after having observed the target word.

These context residuals are not the most intuitive concept, but they turn out to have a very interesting property: they are predictive of MRI data, but not of MEG data. This finding is supported by a range of experiments, including a replication with a new LM (GPT2 instead of ELMO) and with fresh MRI data.

Linguistically, the context residuals are not so easily characterized. The authors refer to them as capturing 'supra-word meaning', motivated with the example of 'Mary finished the apple' as silently assuming that 'Mary ate the apple'. The context residual,

when 'apple' is the target word, would therefore contain semantic information about 'eating'.

The paper represents solid work, reports on a potentially very relevant finding. It does leave open some issues, but I think the contents are sufficient to warant publication now.

In the previous version, the presentation of the paper, and in particular the introduction about its relevance for linguistics and the explanation of the core concepts were suboptimal. This new version now reads very well. I think the introduction is very clear; leaving much of the interpretation of "context residuals" to the discussion works well, for me.

One remaining issue I have with the presentation is that there still is no definition, using a mathematical equation, of 'context residual' in the main text. I would recommend inserting that after line 150. Also, the same concept to is referred to as "residual context embedding" in the main text, and "context residual" in the figure and elsewhere. For readers still trying to get their head around this concept this can be confusing.

A similar issue applies to 'residual word embeddings'. At line 243 no real definition given, and for details readers are referred to Materials & Methods. I think a simple equation could clarify these greatly.

I think these final presentation issues could be left to the authors themselves to decide on.

Author Rebuttal, second revision:

Response to reviewers

Reviewer #2:
Remarks to the Author:
The revisions that the authors have made have, to me, greatly clarified the paper and its main points, and I am now happy with the manuscript as it is.

The clear distinction that the authors now make throughout between supra-word meaning and supra-word embeddings is extremely helpful and makes the overall argument much easier to follow.

I particularly like the new paragraph from lines 49 through 79, which provides a strong motivation for the use of neural language models in studying supra-word meaning, and an insightful statement on what it is that we can or cannot learn from the use of language models. Great point about the usefulness of a single, holistic supra-word embedding as opposed to testing an individual linguistic phenomenon.

I also really like the addition of lines 395-400, where the authors point out why it would not be enough to know that linguistic phenomena X, Y, Z are included in supra-word embeddings, and how the residuals method can be used in the future. I had not understood before that the residual method can in principle be used to *remove* traces of particular linguistic phenomena from the supra-word embedding for more in-depth tests.

This is an important point, I think, and an important contribution to the debate on how we should be using language models in the future.

As I think I already said in my previous rounds of comments, it is good that the authors re-ran their analyses on additional datasets (the movie data, and the radio stories data) and with an additional language model with a different architecture (GPT-2), to show that the null result on MEG data is a true finding, and not a problem with the data.

The code uses Python, a widely used and easily accessible programming language. There are few Python packages required to run the code, and they are all, again, widely used and easily accessible. There was only one additional package I needed to install, and that was installed in less than a minute.

The README is clear and well-organized, and gives the exact commands needed to run, along with expected output. The code itself comes in three short and well organized Python scripts. It has few comments, but the function names are well chosen so one can guess what each function does.

I ran the first three commands suggested in the README. The very first run gave me an error, FileNotFoundError: [Errno 2] No such file or directory: './results/predict_fmri_l_gpt2_with_prev-1-context.npy'. But when I made an empty subdirectory called "results" from the directory with the code, and then re-ran, everything ran without a hitch, and produced exactly the results given in the README.

Concerning the third paragraph of Reviewer 1's comments, I think the authors have addressed this point, which coincides with comments that I made and that also other reviewers made. They have carefully separated the terms "supra-word meaning" and "supra-word embedding", and

they have added, both in their rebuttal and in the main text, a paragraph that I find quite important. It suggests using embeddings to compare to brain patterns exactly because they seem to be based on a mixture of all sorts of different linguistic phenomena

– Katrin Erk

> We appreciate the reviewer's positive feedback about our changes. We also thank the reviewer for the help to improve the paper through the revisions. We will add a "results" directory in our code base to avoid the bug reported by the reviewer.

Reviewer #4:
Remarks to the Author:

I have enjoyed reading the second revision of the paper. As I wrote in my earlier review, I think the paper represents an important contribution to science. There have been a lot of papers in the last few years showing that vectors extracted from deep language models (LMs) can be used to predict brain activation in MRI, EEG and MEG. This paper attempts to make an important advance by disentangling various sources of information.

To this end, the paper defines 'context residuals': the difference between the representation of prior context of a given word that one can directy extract from an LM (before seeing the target word), and the representation of context that one would retroactively expect after having observed the target word.

These context residuals are not the most intuitive concept, but they turn out to have a very interesting property: they are predictive of MRI data, but not of MEG data. This finding is supported by a range of experiments, including a replication with a new LM (GPT2 instead of ELMO) and with fresh MRI data.

Linguistically, the context residuals are not so easily characterized. The authors refer to them as capturing 'supra-word meaning', motivated with the example of 'Mary finished the apple' as silently assuming that 'Mary ate the apple'. The context residual, when 'apple' is the target word, would therefore contain semantic information about 'eating'.

The paper represents solid work, reports on a potentially very relevant finding. It does leave open some issues, but I think the contents are sufficient to warant publication now.

In the previous version, the presentation of the paper, and in particular the introduction about its relevance for linguistics and the explanation of the core concepts were suboptimal. This new version now reads very well. I think the introduction is very clear; leaving much of the interpretation of "context residuals" to the discussion works well, for me.

> We appreciate the reviewer's positive feedback about our changes. We also thank the reviewer for the help to improve the paper through the revisions.

One remaining issue I have with the presentation is that there still is no definition, using a mathematical equation, of 'context residual' in the main text. I would recommend inserting that after line 150. Also, the same concept to is referred to as "residual context embedding" in the

main text, and "context residual" in the figure and elsewhere. For readers still trying to get their head around this concept this can be confusing.

We thank the reviewer for this comment, we added more mathematical detail and an equation in the main text and changed the references in text and figures to all be about residual context.

A similar issue applies to 'residual word embeddings'. At line 243 no real definition given, and for details readers are referred to Materials & Methods. I think a simple equation could clarify these greatly.

We also added more clarification for the residual word embeddings.

I think these final presentation issues could be left to the authors themselves to decide on.

nature research

Final Decision Letter:

Date: 12th October 22 16:04:30
Last Sent: 12th October 22 16:04:30
Triggered By: Kaitlin McCardle
From: kaitlin.mccardle@us.nature.com
To: lwehbe@cmu.edu
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-21-0986C
Message: Dear Dr Wehbe,

We are pleased to inform you that your Article "Combining computational controls with natural text reveals aspects of meaning composition" has now been accepted for publication in Nature Computational Science.

Once your manuscript is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that Nature Computational Science is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see https://www.nature.com/natcomputsci/for-authors). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on

which it is uploaded onto our web site).

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NATCOMPUTSCI-21-0986C and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Computational Science. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

We look forward to publishing your paper.

Best regards,

Kaitlin McCardle
Editor
Nature Computational Science

P.S. Click on the following link if you would like to recommend Nature Computational Science to your librarian: https://www.springernature.com/gp/librarians/recommend-to-your-library

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**