In Press, Psychophysiology

Neural Correlates of Word Representation Vectors in Natural Language Processing Models: Evidence from Representational Similarity Analysis of Event-Related Brain Potentials

Taiqi He[1,2], Megan A. Boudewyn[3], John E. Kiat[4], Kenji Sagae[2], and Steven J. Luck[4,5*]

[1]Center for Neuroscience, University of California, Davis, USA

[2]Department of Linguistics, University of California, Davis, USA

[3]Department of Psychology, University of California, Santa Cruz, USA

[4]Center for Mind & Brain, University of California, Davis, USA

[5]Department of Psychology, University of California, Davis, USA

[*]Corresponding Author

Email: sjluck@ucdavis.edu

Funding: This study was made possible by NIH grants R01MH076226 and R01MH087450 to SJL and a Young Investigator Grant from the Brain & Behavior Research Foundation to MAB.

Abstract

Natural language processing models based on machine learning (*ML-NLP* models) have been developed to solve practical problems, such as interpreting an Internet search query. These models are not intended to reflect human language comprehension mechanisms, and the word representations used by ML-NLP models and human brains might therefore be quite different. However, because ML-NLP models are trained with the same kinds of inputs that humans must process, and they must solve many of the same computational problems as the human brain, ML-NLP models and human brains may end up with similar word representations. To distinguish between these hypotheses, we used representational similarity analysis to compare the representational geometry of word representations in two ML-NLP models with the representational geometry of the human brain, as indexed with event-related potentials (ERPs). Participants listened to stories while the electroencephalogram was recorded. We extracted averaged ERPs for each of the 100 words that occurred most frequently in the stories, and we calculated the similarity of the neural response for each pair of words. We compared this 100×100 similarity matrix to the 100×100 similarity matrix for the word pairs according to two ML-NLP models. We found significant representational similarity between the neural data and each ML-NLP model, beginning within 250 ms of word onset. These results indicate that ML-NLP systems that are designed to solve practical technology problems have a representational geometry that is correlated with that of the human brain, presumably because both are influenced by the structural properties and statistics of language.

1. Introduction

This study addresses the relationship between representations of words in the human brain—as indexed by event-related potentials (ERPs)—and representations of words in machine learning models of natural language processing (*ML-NLP* models). In the Introduction, we begin by describing the general theoretical issue at stake, and then we turn to the nature of word representations in the models. We then describe *representational similarity analysis*, a general method for assessing links between computational models and empirical data (or between different types of empirical data). Because this method has been applied only rarely to ERPs, we provide a general overview of the method. Finally, we describe how the present study was designed to assess links between two specific ML-NLP models and human ERP data.

1.1. How might machine learning models of natural language processing be related to the human brain?

As machine learning models grow more competent at the human task of natural language processing, it is natural to question whether these ML-NLP models exhibit any functional resemblance to the neural systems that underlie human language processing. Machine learning models of natural language processing might be expected to operate very differently from the human brain because they run on very different hardware architectures (e.g., silicon chips capable of trillions of floating-point operations per second). They are not constrained by the relatively slow and stochastic firing of individual neurons, but they do not benefit from the massive parallelization that characterizes the architecture of the brain. Moreover, ML-NLP models have "evolved" rapidly through human engineering to solve a set of specific practical problems rather than evolving gradually through natural selection to maximize the overall fitness of the organism.

However, natural language production and understanding is such a complex problem that there may be few optimal solutions, and both natural selection and engineer-guided evolution may therefore converge on the same solution. Moreover, if an ML-NLP model and a human brain are trained on similar inputs (i.e., examples of the same natural language), the structural and statistical properties of linguistic input itself might lead to similar representations in these two systems. Thus, the representations used by ML-NLP models and human brains might be unrelated (owing to their different architectures and goals), highly similar (owing to similarity in their training), or somewhere in between. The goal of the present study was to identify the extent of convergence between ML-NLP models and the human brain in the specific case of lexical representations.

ML-NLP models have been created that can accomplish various practical tasks such as question answering and named entity extraction (Devlin et al., 2018; Peters et al., 2018). In particular, we examined ML-NLP models of *word embeddings* or *word vectors*. Word embeddings are word representations that quantitatively capture the notion that words that appear in similar sentential environments (in terms of the surrounding words) tend to have similar meanings, an idea dating back to the *distributional hypothesis* (Firth, 1957; Harris, 1954). For example, there is a limited set of words that are likely to appear in the context "The fluffy brown _____ groomed itself on the windowsill."

Building on this idea, computational linguists have trained models that use contextual information to map words to continuous vectors, representing each word as a point in a high-dimensional space (see Figure 1A). These vectors or word embeddings can then be used to determine the similarity between any given pair of words (i.e., the distance between them in the

high-dimensional space), and the vectors can be used to improve the performance of a variety of downstream tasks. For example, "bucket" and "pail" have similar word embedding vectors, and this similarity can be used to ensure that an Internet search request that includes the word "bucket" will also return results that contain the word "pail".

Word embeddings are also interesting to cognitive psychologists because they loosely resemble spreading activation theories insofar as semantically and syntactically similar words are represented close together in the embedding space (Foltz, 1996; Foltz et al., 1998; Landauer et al., 1998). However, the individual dimensions used by ML-NLP models do not typically have an obvious interpretation and are instead statistical abstractions.

1.2. Representational similarity analysis

Comparing computational models such as these with the human brain can be challenging because superficial differences between them may obscure fundamental similarities. In particular, it may be difficult to map the representational elements in a model (e.g., a vector in a word embedding space) onto the macro-level population measures of brain activity that can be obtained noninvasively from human research participants. That is, even if there is a perfect relationship between the representation of a word in a model and the representation of that word in the brain, we may not be able to detect this relationship because we do not know the exact function that maps the model's representation of the word onto the pattern of brain activity elicited by that word, especially as measured macroscopically by the pattern of BOLD activity across voxels in fMRI or the pattern of voltages across electrode sites in EEG.

This problem can be addressed by means of representational similarity analysis (RSA), which overcomes superficial differences between representational systems by comparing their similarity structures. In RSA, the structure of a given representational system (the *representational geometry*) is quantified by feeding a large number of different inputs into the system, measuring the pattern of activity elicited by each input, and computing the similarity of the response to each possible pair of inputs. This is illustrated in Figure 1, which shows how RSA could be applied to the representations of a set of 100 different words.

In this example, we feed 100 different words into a ML-NLP model, obtain a vector for each word in the word embedding space, and calculate the similarity between each pair of word vectors in this space. The similarity between two words is defined in terms of the angle between the vectors in the embedding space[1]. In Figure 1A, for example, the words "see" and "hear" are relatively close together in the embedding space and would have a high similarity score, whereas "quick" is far away and would have low similarity scores with respect to "see" and "hear." For the set of 100 words in this example, we would form a 100×100 *representational similarity matrix* that contains all the pairwise similarity scores, as shown in Figure 1B. This matrix captures the representational geometry of the model in a manner that is abstracted away from the original dimensions of the embedding space.

We can present the same set of 100 words to a human research participant and record the participant's neural responses for each word. We can then quantify the similarity between the neural responses to each pair of words. There are several potential metrics of similarity, but many studies simply use the correlation between the neural patterns. For example, we could obtain the ERP scalp distribution for each word during a given time period and compute the

[1] Note that only three dimensions are shown for the word embedding space in Figure 1A, but ML-NLP models typically use hundreds of dimensions.

correlation[2] between the scalp distributions elicited by two words, as shown in Figure 1C. If we obtain the correlation between each pair in our set of 100 words, we can construct a 100×100 representational similarity matrix for the neural responses, as shown in Figure 1D. This matrix captures the representational geometry of the neural response in a manner that is abstracted away from the original dimensions of the neural recording (which might be electrode sites, time points, voxels, etc., depending on the recording method). No matter what the original dimensions are, we can form a 100×100 representational similarity matrix from the correlations between the patterns of neural activity elicited by each word pair.

By creating 100×100 representational similarity matrices for both the ML-NLP model and the neural activity, we have quantified the representational geometry of each system in the same units (a 100×100 matrix of similarity values). We can then ask whether the model and the brain have similar representational geometries by simply calculating the correlation between the two matrices. For reasons discussed in the Method section, a rank-order correlation is used for comparing the two matrices. The upper and lower triangles of the matrices are mirror symmetrical, and the values are always 1 along the diagonal, so the correlations between the matrices are computed from the lower triangles (or from the upper triangles, which yields identical results).

This approach has been widely used to compare neural network models of visual image classification (deep convolutional neural networks) with human neural responses to images. For example, Cichy et al. (2016) obtained MEG and fMRI responses to a set of 188 natural images in a set of human research participants, and they also fed these images into a neural network model that had been trained to classify objects. A separate representational similarity matrix was obtained for each area of visual cortex in the fMRI data, for each time point in the MEG data, and for each of the 8 layers of the model. The matrices from the lower layers of the model showed the highest correlations with the matrices for the early stages of the visual processing pathway in the fMRI data and with the matrices for the early time points in the MEG data. By contrast, the matrices from the higher layers of the model showed the highest correlations with higher-level brain regions in the fMRI data and later time points in the MEG data. This is exactly what would be expected given that the flow of information from lower to higher levels of the model should map onto the flow of information from lower to higher brain areas in the fMRI data and the flow of information over time in the MEG data. A similar correspondence between layers and time points was observed between a convolutional neural network model of scene categorization and the ERP responses to a set of scenes (Greene & Hansen, 2018). We have also used this approach to link ERPs with models of the spatial distribution of saliency and semantic informativeness in visual scenes (Kiat et al., under review). These results provide an important proof of principle for using RSA to link computational models to ERP measures of brain activity.

1.3. The present study

The present study compared ERP data to two models of word embeddings that have quite different architectures but that both aim to represent discrete words as continuous vectors. The first, *Fasttext* (Bojanowski et al., 2017), is an extension of the *word2vec* approach (Mikolov et al., 2013), which involves correlating a word and its contexts in a corpus. It was trained with text from English Wikipedia to predict the surrounding words in an utterance given an input word.

[2] The Pearson *r* correlation coefficient has some useful properties as a measure of similarity. One is that it is simple and well-understood. Another is that it represents the similarity of two neural patterns irrespective of the amplitude of the neural response. For a detailed comparison of distance metrics, see Guggenmos et al. (2018).

The model also takes into account an approximation of the morphological similarity between words by giving embeddings to within-word character sequences. For example, the word "language" contains the following 3-character sequences: <la, lan, ang, ngu, gua, uag, age, ge>, with the angle brackets representing word boundaries. The vectors of the subword sequences are summed to enhance the representation of a word. Consequently, if two words are similar in spelling, they will have similar vectors by virtue of having overlapping subword sequences. For example, the vector for "language" will be very similar to the vector for "languages" (or even misspellings of "language"), and this makes the model more robust.

The second model of word embeddings we will consider, *ELMo* (Peters et al., 2018), is based on a recurrent neural network trained on 800 million words from web-crawled news articles in English (Chelba et al., 2014). The embeddings drawn from this model are context dependent, meaning that words that appear in different environments will produce different representations. For example, the word "orange" would have a different embedding when preceded by "the color" than when preceded by "the juicy." This reflects the structure of English, because a given word token can have different meanings in different contexts. By contrast, Fasttext has a single representation of "orange" that is similar to both the representations of other colors and the representations of other fruits. Note that multiple factors impact word embeddings, including predictability as well as meaning.

We chose the Fasttext and ELMo models because they are structurally very different (a simple statistical representation versus a recurrent neural network) but are both based solely on statistical regularities in the sequence of words and contain no explicit semantic information. Thus, it is plausible that the pattern of representational similarity in these models will be quite different from the pattern in the human brain, which presumably has explicit representations of concepts such as "see," "hear," and "quick." On the other hand, the pattern of word representations in the human brain may be strongly influenced by statistical regularities in the linguistic input (Saffran et al., 1996), and this might lead the representational similarity structure of the brain to be to very similar to the representational similarity structure of the ML-NLP models.

The conventional way of measuring similarity between word embeddings in the models is to take the cosine distance between them (which is nearly identical to the correlation between them[3]), and this is how we constructed the representational similarity matrices for the models. The most intuitive way of measuring the similarity of words in the human mind would be to ask people to explicitly rate the similarity of each pair of words (as in the WordSimilarity-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2014) databases). However, collecting human data this way is slow and expensive, especially considering that a 100-word similarity matrix requires judgments of 4950 unique pairs of words. In addition, explicit ratings of similarity are unidimensional, and very different ratings of similarity may be obtained depending on how people collapse the underlying multidimensional space into a unidimensional rating scale. Moreover, an explicit similarity judgement task may not be well suited for classes of words with little semantic content (e.g., function words such as "the" and "for").

Because of these shortcomings of explicit similarity judgments, we assessed the similarity of word representations in humans by means of ERPs, which are inherently multidimensional and can be obtained during natural language comprehension without the need

[3] The cosine distance quantifies the difference in the angles of the two vectors, disregarding differences in the lengths of the vectors. This is analogous to the fact that the Pearson *r* correlation coefficient quantifies the similarity of two patterns while disregarding the overall amplitudes of the patterns.

for an explicit similarity rating. Our ML-NLP models were designed for natural language, so we chose an ERP data set from a prior experiment in which participants simply listened to two half-hour Sherlock Holmes stories that were spoken with normal timing and inflection. An important strength of this data set is that participants were engaged in completely natural language processing, including both the nature of the linguistic input and the processing that was applied to this input. There is perhaps nothing more natural for humans than to listen to stories. Moreover, RSA often benefits from the lack of an explicit task, because task-induced categorization processes may influence the similarity structure of the ERPs (e.g., due to nonlinguistic activity such as the P3 wave).

The downside of this type of data set is that we did not exert experimental control over the number of words, the types of words, or the order of words. As a result, some words occurred frequently, allowing for clean averaged ERPs, but other words occurred rarely. In addition, the current word was often predictable from the preceding words. However, our ML-NLP models were trained by predicting words from their contexts in natural language, so this data set matches the models quite well. In addition, RSA is particularly well suited for studying natural stimuli, including natural speech and natural visual scenes. Thus, the strengths of this data set outweighed the weaknesses for the goals of the present study.

To deal with the varying numbers of instances per word, our main analyses focused on the 100 words that occurred most commonly in these stories (see Table 1, which lists the words and the number of occurrences of each word). An averaged ERP waveform was constructed for each of these words. When we assessed the similarity in neural activity for each pair of words, we took into account both the shape of the waveform at each site and the distribution of voltage over the scalp[4] (see Figure 2 and Method for details). This gave us a 100×100 similarity matrix, which we could then compare with the similarity matrices from the Fasttext and ELMo models.

The representational similarity between the neural data and a given model was quantified as the rank-order correlation between the corresponding representational similarity matrices. Semipartial correlations were used to assess the extent to which each ML-NLP matrix explained unique variance in the neural representational similarity matrix. We also computed ERP similarity matrices at each individual time point in the ERP waveforms, focusing solely on the similarity of the scalp distributions at a given time point, which made it possible to examine the time course of the relationship between the ERP data and the models.

If the ML-NLP models examined here represent words very differently than the human brain because of their different origins and goals, then the representational similarity matrices of the models should be largely unrelated to the representational similarity matrix derived from the neural data. By contrast, if the models and the human brain represent words similarly because they were both trained on the same language and have converged on similar solutions to the problems posed by natural language processing, then the similarity matrices of the models should be correlated with the similarity matrix from the neural data.

We can use the same approach to address whether Fasttext and ELMo have converged on similar representational structures (i.e., by computing the correlation between the similarity matrices of the two models). The strength of the correlation between these two models can be used as a benchmark for assessing the strength of the correlation between the neural data and each model. That is, given that Fasttext and ELMo were both designed to achieve the same

[4] This combined use of waveform shape and scalp distribution gave us an overall metric of the similarity in the neural representations of each pair of words, which we could compare with the overall similarity of the ML-NLP embeddings for each pair of words.

fundamental goal, the magnitude of the correlation between the Fasttext and ELMo similarity matrices provides a reasonable benchmark for the highest correlation that might be expected for the correlation between the ERP-based similarity matrix and each of the ML-NLP models.

2. Method

2.1. Experimental Paradigm and EEG Recordings

This study made use of EEG data collected by Boudewyn and Carter (2018), and a more detailed description of the participants, experimental paradigm, and recording procedures can be found in their paper. The data and our analysis scripts are publicly available at https://osf.io/zft6e/.

Complete data sets with an acceptable number of artifact-free EEG segments were available from 38 participants (11 male and 27 female). Two participants were excluded because they did not complete the experiment and two more were excluded because of excessive artifacts (see below). The participants were right-handed, native English speakers enrolled as students at the University of California, Davis (mean age 20.4 years).

The participants simply listened to two Sherlock Holmes stories (*The Three Students* (Doyle, 1905), 34.4 min duration, and *The Emerald Crown* (Doyle, 1992), 38.4 min duration), narrated by a woman with typical American English inflections and at a natural speaking rate. There was no explicit task. The original study was designed to assess how lapses of attention impacted the ERPs elicited by a subset of the words. Therefore, on 54 occasions, the stories were interrupted by questions asking the participants about their current attentional state. The EEG and behavioral data from these questions were not analyzed in the present paper.

The EEG was recorded from 29 scalp channels, with the right mastoid being used as the recording reference. Bipolar vertical and horizontal electrooculogram recordings were obtained using electrodes above, below, to the right, and to the left of the eyes. The signals were amplified with half-amplitude bandpass cutoffs at 0.05 and 100 Hz, digitized at 500 Hz, and later down-sampled to 250 Hz.

2.2. EEG preprocessing

Data preprocessing was performed using Matlab with the EEGLAB toolbox (Delorme & Makeig, 2004) and the ERPLAB plugin (Lopez-Calderon & Luck, 2014). The continuous signals were filtered offline using noncausal Butterworth filters (high-pass: half-amplitude cutoff at 0.1 Hz, 12 dB/oct roll-off; low-pass: half-amplitude cutoff at 20 Hz, 48 dB/oct roll-off). Independent component analysis (ICA) was used to correct for eyeblinks.

Event codes were inserted into the EEG files at the estimated onset time of each word on the basis of a procedure that made use of the transcripts of the two stories and the audio recordings. First, the transcripts were tokenized with the NLTK tokenizer (Bird et al., 2009). The onset time of each token was then estimated using the Montreal Forced Aligner (McAuliffe et al., 2017), and the event code was inserted at the corresponding time within the EEG data file. The results were compared with a sample of words that were hand-labeled by one of the authors (M.A.B.). The mean absolute error of the aligner relative to the expert experimenter was 18.73 ms for *The Emerald Crown* and 20.32 ms for *The Three Students*. Additional details are provided in online supplementary materials. Note that any random error in the timing of the event codes will have an effect equivalent to applying a low-pass filter in the time domain (Luck, 2014). These timing errors may slightly "smear out" the data, but this should mainly impact the short-

latency ERP components and not the longer-latency components that are most likely to contain information that is related to the word embeddings in the ML-NLP models. Thus, any errors in event code timing should have relatively little impact on the primary goal of the study, which is to assess the extent to which word representations in ML-NLP models are similar to those in the human brain.

Data analysis was limited to the 100 words that occurred most frequently across the two stories (see Table 1 for the actual words, along with the mean number of occurrences of each word both before and after artifact rejection). RSA benefits from having a large number of individual stimuli. This increases the size of each representational similarity matrix, which in turn increases the degrees of freedom in the correlation between a pair of matrices. It also increases the likelihood that the set of stimuli is reasonably representative. However, RSA also benefits from a high signal-to-noise ratio in the ERPs for each stimulus, which can be achieved by averaging together a large number of trials for each stimulus. To balance these two competing considerations, we chose the 100 words that occurred most often in the two stories for the primary analyses. This gave us a reasonably representative set of words and 4948 degrees of freedom when correlating a given pair of representational similarity matrices. It also gave us a minimum of 12 instances of each word and therefore reasonably clean averaged ERPs. We also provide a secondary analysis of all 14,804 word tokens that appeared in the two stories.

The EEG data for the 100 selected words were epoched, using the event codes at the estimated word onset as time zero, with a 200 ms prestimulus baseline and a 1000 ms period after the word onset. The epochs were baseline corrected using the mean voltage during the prestimulus period. A moving window peak-to-peak artifact rejection algorithm was applied (with a threshold of 200 μV) to eliminate epochs with implausibly large voltage excursions. This led to a mean rejection rate of 6.34% (range: 0-50.7%**)**. We always exclude participants for whom more than 25% of trials are rejected (Luck, 2014), and 2 participants were excluded for this reason, yielding a final sample of 34 participants.

All artifact-free occurrences of a given word were averaged together to yield a separate ERP waveform for each word. Different forms of the same word (e.g., “hit” and “hits”) were treated as different words. Contractions were divided into separate words (e.g., “didn’t” was separated into “did” and “n’t”). Note that the averaged ERP for a given word represents the average response to the word across the different contexts in which the word occurred. As described in the next section, the ML-NLP word embeddings used for the main analyses were also constant across word contexts.

2.3. Word Embedding Models

Two word embedding models were used. The Fasttext model (Bojanowski et al., 2017) uses an enhancement of the skip-gram approach introduced by Mikolov et al. (2013), adding subword n-grams so that spellings contribute to similarity and to allow generalization to untrained words. We simply obtained the Fasttext embedding vector for each of the 100 words shown in Table 1, combining the subword sequences.

Table 1. Top 100 words (mean number of instances after artifact rejection, standard deviation, true frequency).

Function words					
and (360.24, 21.71, 376)	the (317.56, 17.36, 331)	i (277.94, 16.09, 289)	that (278.12, 18.14, 291)	to (244.56, 16.28, 256)	was (237.12, 13.36, 247)
you (208.24, 14.82, 218)	he (192.91, 12.35, 202)	his (178.59, 10.87, 187)	had (152.24, 9.88, 159)	have (152.71, 9.96, 159)	of (143.21, 7.76, 149)
it (139.53, 8.77, 146)	is (133.68, 8.12, 140)	my (133.94, 8.37, 140)	not (115.47, 8.03, 121)	with (112.00, 6.68, 117)	in (104.56, 6.25, 109)
as (97.74, 7.82, 103)	me (94.09, 7.40, 99)	for (98.62, 7.47, 104)	your (95.35, 6.52, 100)	this (93.74, 5.78, 98)	be (89.71, 5.85, 94)
but (85.03, 5.93, 89)	there (77.09, 5.37, 81)	no (77.29, 3.87, 80)	which (71.09, 5.79, 75)	him (69.44, 5.56, 73)	on (69.65, 5.04, 73)
so (68.91, 4.52, 72)	what (67.47, 4.84, 71)	would (65.06, 4.03, 68)	then (63.18, 4.09, 66)	when (61.06, 3.92, 64)	could (63.09, 3.87, 66)
been (58.82, 3.53, 61)	very (58.79, 3.11, 61)	she (55.12, 4.84, 58)	at (55.38, 4.08, 58)	by (53.65, 3.48, 56)	were (54.35, 4.07, 57)
from (53.38, 3.54, 56)	all (53.53, 3.79, 56)	up (47.47, 3.52, 50)	now (46.71, 3.37, 49)	do (52.24, 4.27, 55)	can (43.12, 2.86, 45)
into (42.15, 2.89, 44)	out (42.91, 3.15, 45)	who (40.29, 2.79, 42)	if (39.79, 3.06, 42)	did (37.21, 2.91, 39)	we (36.56, 2.43, 38)
down (35.65, 2.09, 37)	only (35.53, 2.34, 37)	must (34.53, 2.34, 36)	will (36.03, 2.71, 38)	has (33.74, 2.12, 35)	a (32.03, 1.24, 33)
about (30.44, 2.83, 32)	where (30.74, 2.09, 32)	our (30.41, 2.70, 32)	over (28.74, 1.91, 30)	them (28.79, 1.81, 30)	her (28.59, 2.19, 30)
am (27.76, 2.03, 29)	well (26.82, 2.11, 28)	should (26.88, 1.77, 28)	they (24.56, 2.39, 26)	an (15.26, 1.42, 16)	n't (14.24, 1.46, 15)
are (11.38, 0.95, 12)					
Content words					
holmes (190.74, 11.56, 200)	said (163.18, 9.48, 171)	mr (109.71, 6.30, 115)	one (79.26, 5.14, 83)	asked (72.59, 4.78, 76)	room (61.71, 2.98, 64)
man (59.24, 4.05, 62)	payton (56.82, 3.53, 59)	bannister (51.12, 3.44, 53)	holder (47.82, 3.93, 51)	think (45.12, 3.09, 47)	see (44.21, 2.29, 46)
answered (43.88, 3.36, 46)	door (42.85, 2.70, 45)	sir (42.38, 2.45, 44)	back (42.18, 2.75, 44)	left (32.88, 1.63, 34)	came (31.65, 2.17, 33)
table (31.12, 1.72, 32)	papers (29.65, 2.98, 31)	three (29.88, 1.61, 31)	went (30.38, 2.47, 32)	matter (28.68, 2.17, 30)	window (29.65, 1.89, 31)
little (26.59, 2.08, 28)	other (26.56, 2.48, 28)	come (26.65, 2.13, 28)			

The ELMo model (Peters et al., 2018) uses a deep recurrent neural network language model that returns token-by-token embedding vectors given a sentence as its input. Consequently, whereas Fasttext produces a static, context-independent embedding vector for a given word, the embedding vector returned by ELMo for a given token depends on the context in which that word occurs. Because the ERP for a given word in the present study was created by

averaging across all the contexts in which the word appeared in the two stories, we computed an average ELMo embedding vector for each word in the contexts of these two stories. That is, the ELMo embeddings were obtained by feeding each sentence from the two stories into the mode and then averaging the embeddings over all occurrences of each of the 100 words shown in Table 1. A secondary analysis was conducted for ELMo in which the context-specific embeddings were used.

2.4. Representational Similarity Analysis

RSA was performed using custom Python scripts, which are available at https://osf.io/zft6e/. For both Fasttext and ELMo, we quantified the similarity between each pair of words as the cosine similarity between the word embedding vectors (because that is the most common metric of similarity in studies of word embedding spaces). The cosine similarity is the dot product of the two vectors after normalization, which is equivalent to the cosine of the angle between the two vectors. If all vectors were normalized to the same length, cosine similarity would yield the same rank order as the Euclidean distance. This procedure yielded a 100×100 matrix representing the pairwise similarities between the individual words for each of the two models. Note that the same matrices were used for each participant and each time point.

We constructed a corresponding similarity matrix for the ERP data, separately for each participant. The procedure is illustrated graphically in Figure 2. We first subtracted the average ERP across words (with each word given equal weight) from each individual-word ERP. This eliminates any correlation between the ERPs for different words resulting from ERP activity that is constant across words. The waveforms during the poststimulus period from each channel for a given word were then concatenated end-to-end, producing a single vector of voltage values that includes both the spatial and temporal variation in voltage corresponding to that word. The similarity between each pair of words was quantified as the Pearson *r* correlation between the vectors of voltage values for the two words. This yielded a single similarity value for each word pair, which were organized into a 100×100 matrix using the same arbitrary ordering used for the Fasttext and ELMo similarity matrices.

The upper and lower triangles of this similarity matrix are mirror images of each other, and the diagonal separating them represents the similarity of a word with itself. When comparing matrices, we excluded the diagonal and the lower triangle. Given that relationship between a given pair of matrices might not be linear, we used the Spearman rho rank order correlation coefficient to quantify the similarity between the ERP similarity matrix for a given participant and the Fasttext or ELMo similarity matrix (which were identical across participants). We also used semipartial rank order correlations to quantify the amount of variance in the ERP similarity matrix that could be uniquely explained by each of the two ML-NLP models.

Because our main dependent variability was the Spearman rho correlation coefficient (i.e., our metric of representational similarity), and correlation coefficients are unlikely to be normally distributed, we used nonparametric statistical tests. First, we used bootstrapping with 100,000 random draws to compute 95% confidence intervals for the mean Spearman rho across participants. Second, we used the nonparametric Wilcoxon signed-rank test to determine whether the mean Spearman rho was significantly greater than zero. One-tailed tests were used when comparing rho values against chance because negative values are typically uninterpretable in RSA. We used an alpha of .05 for all statistical analyses.

To make sense of the range of similarity values, we calculated the noise ceiling for the ERP data, which provides an estimate of the highest correlation that could be obtained given the

noise in the data (Nili et al., 2014). For analyses using Spearman rho, the upper bound is the averaged rho of the correlations between individual correlation matrices and the grand average correlation matrix. The lower bound is the same, except using a separate grand average correlation matrix for each participant that excludes that participant's data.

We also computed a grand average ERP similarity matrix by averaging the ERP similarity matrices across participants and examined the Spearman rho correlation between this matrix and the similarity matrix for each of the ML-NLP models. This provides an estimate of the magnitude of the correlation between the "average" person (with minimal noise) and each of the ML-NLP models. We used bootstrapping with 100,000 random draws to compute the 95% confidence interval for each of these correlations. Note, however, that averaging the similarity matrices across participants can be problematic (Ashby, 2019; Ashby et al., 1994), so these analyses should be treated with caution.

2.5. Time Course Analyses

To obtain temporal information, we also performed the RSA separately at each sample point. That is, rather than concatenating each entire ERP waveform for the various electrode sites, we quantified the similarity between two words as the similarity (Pearson *r* correlation) between the scalp distributions observed for the two words at a given time point. Each time point was treated separately. The ERP similarity matrix at a given time point was then compared to the representational similarity matrix for each of the two NLP models using the Spearman rho correlation coefficient. This analysis was performed separately for each subject.

The noise ceiling was estimated separately at each time point, using the same approach as in the whole-waveform analyses.

We used a one-tailed Wilcoxon signed-rank test to determine whether the mean Spearman rho across participants was significantly greater than zero at each time point. Given the large number of individual time points, we applied a false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) to the results from each model to correct for multiple comparisons.

3. Results

3.1. Basic characteristics of the words and the ERP waveforms

Table 1 shows the 100 most frequent words in the two stories, along with the number of instances of each word across the two stories. Because some trials were excluded from the ERP averages due to EEG artifacts, the number of words per averaged ERP differed across participants. Table 1 therefore shows the mean number of instances of each word after artifact rejection, the standard deviation of the number of words, and the number of instances in the stories prior to rejection. We classified each word as a function word (e.g., "and", "the") or a content word (e.g., "room", "came"), because these two categories of words are known to elicit different ERPs when presented in isolation (Brown et al., 1999; Kutas & Hillyard, 1983). Of the 100 words, 73 were function words and 27 were content words. Most of our analyses disregarded this classification, but secondary analyses were conducted separately for function and content words.

Figure 3 shows the grand average ERP waveforms separately for the average of the 27 content words and for the average of the 73 function words at the midline electrode sites. The ERP waveforms were quite different for these two classes of words. For example, the voltage in

the N400 latency range was more negative for the content words than for the function words, as in previous research (Neville et al., 1992). Differences between the waveforms were also observed at or before the onset of the word, suggesting that at least some of the differences reflect differential overlap from preceding words. An inevitable consequence of examining ERPs during natural language comprehension is that the voltage at a given time point is impacted by the previous words as well as the current word. However, this is not an artifact of the ERP method per se: The cognitive processing of one word presumably does not stop when the next word is heard.

3.2. Similarity between the NLP models

Figure 4 shows the similarity matrices for Fasttext and ELMo models, along with the similarity matrix for the ERP data (averaged across participants, and based on the entire spatiotemporal pattern of the ERPs). The words are divided into function and content words (separated by a white line), and words within each group are ordered according to the number of occurrences of each word in the two stories. A close inspection of the matrices for the two ML-NLP models reveals some similarities in the patterns. For example, the similarity matrices for both models contain a bright yellow vertical band (indicating high similarity between word pairs) and then a dark purple vertical band (indicating low similarity) just to the right of the line separating the function and content words (see purple arrows in Figure 4). However, given that both models attempt to represent word embeddings as continuous vectors on the basis of the surrounding words, their similarity matrices were far from identical. Indeed, the Spearman rho correlation between them was only 0.49 ($p<.001$)[5].

The grand average ERP similarity matrix shown in Figure 4 also bears some resemblance to the similarity matrices for the two NLP models (see, e.g., the green boxes). The Spearman rho correlation between the ERP matrix and the ELMo-ERP matrix was 0.30 ($p<.001$), and the correlation between the ERP matrix and the Fasttext matrix was 0.21 ($p<.001$). Thus, the representational similarity between the group-level ERP data and each of the ML-NLP models was only moderately lower than the similarity between the two models.

However, averaging similarity matrices across individuals can be problematic (Ashby, 2019; Ashby et al., 1994), so these correlations should be treated with caution. An example of the types of incorrect conclusions that can result from averaging the similarity matrices across participants is provided in the online supplementary materials. All subsequent analyses used single-participant ERP similarity matrices.

3.3. Single-participant RSA findings

Our primary analyses focused on comparing single-participant ERP similarity matrices to the matrices for the two models. As shown in Figure 5, both the ELMo-ERP and Fasttext-ERP correlations were above zero for every participant. The mean ELMo-ERP correlation was 0.086 (95% CI [0.07 0.10]), and the mean Fasttext-ERP correlation was 0.057 (95% CI [0.04 0.07]). A sign test indicated that each of these was significantly greater than zero ($p<.001$).

[5] This analysis was limited to the 100 words that were used in the main comparison with the ERP data. We also computed the representational similarity between the two models using all 2339 distinct words that were present in the two stories. This actually led to a lower correlation (0.38, $p <.001$) between the two models than the correlation of 0.49 observed with only the top 100 words. We also trained Fasttext on the same dataset that ELMo was trained on, and the correlation between the matrices for the two models (using the original set of 100 words) increased only slightly from 0.49 to 0.51.

The single-participant data contained substantial noise, so it is not surprising that these correlations were much lower than those obtained for the grand average ERP similarity matrix (and lower than the kinds of correlations that are often observed in psychological research)[6]. To put these correlations into context, we computed the *noise ceiling*, which indicates the best correlation that could be expected given the noise in the data (Nili et al., 2014). The lower and upper bounds of the noise ceiling were 0.15 and 0.26, respectively. Thus, we estimate that the ELMo accounted for between 33.3% and 56.0% of the explainable data, and Fasttext accounted for between 22.1% and 37.2%.

A permutation test was applied (by randomly shuffling the labels of the embeddings) to test whether the correlation between each of the two models and the ERP data was statistically significant for each individual participant. All 34 participants exhibited positive Fasttext-ERP and ELMo-ERP correlations, which were statistically significantly above chance for each participant in a permutation test (even when applying a false discovery rate correction (Benjamini & Hochberg, 1995)). Thus, although the correlations were low by conventional standards, they were remarkably consistent across participants.

Figure 5 also shows the semipartial correlations for the single-participant ERP data, which indicate the extent to which the ERP similarity matrix was correlated with a given ML-NLP matrix after partialling out the correlation with the other NLP matrix. The mean ELMo-ERP semipartial correlation across participants (after partialling out Fasttext) was well above zero (mean = 0.056, 95% CI [0.045 0.067], sign test $z = 5.49$, $p < .001$), whereas the mean Fasttext-ERP semipartial correlation (after partialling out ELMo) was only slightly and nonsignificantly greater than zero (mean = 0.014, 95% CI [0.006 0.022], sign test $z = 1.37$, $p = .085$). Thus, the representational similarity matrix for ELMo explained significant unique variance in the single-participant ERP representational similarity matrices.

We also asked how well the two ML-NLP models together could account for the ERP data. This was quantified by putting the similarity matrices for both models into a single regression model and computing the multiple *r* for predicting the ERP similarity matrix (separately for each participant). We found a multiple *r* of 0.094, which was slightly higher than the correlations for each model separately.

3.4. Separate analyses of content and function words

One possible criticism of the observed RSA effects is that they may merely reflect overall differences in ERPs between content and function words and not the representational structure within these coarse categories. If this were true, then the representational similarity between the models and words within each category should be zero. We tested this hypothesis by conducting separate RSA analyses (using single-participant ERP similarity matrices) for the 27 content words and for the 73 function words. Pronouns are somewhat different from other function words, so we also conducted an analysis with the 59 function words that excluded the pronouns. Table 2 shows these correlations averaged across participants. Although these model-ERP correlations were lower than the correlations observed for the entire set of 100 words, they were significantly greater than zero (see Table 2). These findings demonstrate that the results for the entire set of words was not driven by the simple distinction between function and content words.

[6] It would also be possible to reduce the noise by obtaining a grand average ERP waveform across participants for each word and then creating the ERP similarity matrix from these grand average waveforms. However, this assumes that the patterns of scalp distributions for the different words are identical across participants, which is unlikely to be true given the variability of cortical folding patterns across individuals.

Table 2. Representational similarity correlations computed separately for content and function words.

	Fasttext	ELMo
Content words (noise ceiling bounds [0.11 0.22])	0.033 (95% CI [0.012 0.054], p < 0.01)	0.048 (95% CI [0.025 0.071], p < 0.01)
Functions words (noise ceiling bounds [0.18 0.27])	0.060 (95% CI [0.046 0.074], p < 0.01)	0.086 (95% CI [0.070 0.103], p < 0.01)
Functions words excluding pronouns (noise ceiling bounds [0.15 0.25])	0.072 (95% CI [0.054 0.089], p < 0.01)	0.074 (95% CI [0.057 0.091], p < 0.01)

3.5. Equalizing the number of instances of each word

Another concern is that word-to-word differences in the number of EEG epochs being averaged together could be a potential confound. In general, if a word appears in the dataset more frequently, the ERP waveform for that word would be less noisy, and this might artificially create stronger correlations between the ERPs to frequently occurring words. To address this possibility, we conducted analyses in which we subsampled from the available trials to equate the number of trials in each averaged ERP. We selected words that appeared at least 30 times across all participants after artifact rejection (68 of the 100 words) and randomly sampled 30 trials for each word without replacement before creating the averaged ERPs and performing the representational similarity analysis (using single-participant ERP similarity matrices). We repeated the analysis 100 times (with different random samplings of 30 instances of each word) and averaged the rho values across repetitions. The mean correlation with ELMo was 0.030, and the averaged mean correlation with Fasttext was 0.024. These correlations are substantially smaller than those obtained in the original analyses, but this would be expected because the smaller number of trials led to noisier data. Indeed, the noise ceiling was substantially lower ([0.022 0.173]). The mean correlations were actually higher than the noise ceiling, suggesting that the reduction in rho values was mainly due to the increased noise in the data. Therefore, frequency of occurrence was likely not a major contributor to the correlation between the ERPs and the word embeddings.

3.6. Single-trial analyses

Averaging across instances of a given word is useful for increasing the signal-to-noise ratio of the ERP data, which in turn increases the possible range of correlations (the noise ceiling). However, it is possible to perform RSA on single-trial data if one is willing to accept a much lower ceiling on the range of possible correlations. We therefore conducted a single-trial analysis of the present data.

The analysis was identical to the main single-participant analysis of the top 100 words, except that each instance of a word was treated as a completely separate stimulus. That is, to construct the ERP similarity matrix, we obtained the correlation between the single-trial EEG epochs for each pair of stimuli, disregarding whether two stimuli were different words or different instances of the same word. There were 14804 different word tokens, so this could potentially create a 14804×14804 ERP similarity matrix for each participant. However, EEG epochs containing artifacts were excluded, separately for each participant. A mean of 13982 epochs was present after artifact rejection, so the average EEG similarity matrix was 13982×13982.

A similarity matrix with the same dimensionality was created for each of the two ML-NLP models, separately for each participant (to account for the differences in which epochs were rejected for different participants). When two different cells coded the similarity between the same two words, the similarity value was simply repeated in these two cells. We then computed the Spearman rho correlation between the single-trial EEG similarity matrix and the corresponding ML-NLP similarity matrices to quantify representational similarity.

The representational similarity values obtained from this single-trial approach were extremely small but were statistically significant for both Fasttext (mean = 0.0007, 95% CI [0.0005 0.0009], sign test $z = 4.80$, $p < .001$) and ELMo (mean = 0.0015, 95% CI [0.0012 0.0018], sign test $z = 5.83$, $p < .001$). The fact that they were orders of magnitude smaller than the values obtained for the averaged ERP data is not surprising given the greater noise level of the single-trial data. Unfortunately, our approach to computing the noise ceiling could not be used for this analysis because the same set of trials was not used for each participant (because of artifact rejection). However, the fact that the representational similarity values were significantly greater than zero, and that the relative ordering of the values for Fasttext and ELMo was the same as in the main analysis, demonstrates the feasibility of using this approach.

Whereas word embeddings for Fasttext are independent of the context in which a word appears, ELMo generates a separate embedding for a given word depending on its context. In the analyses so far, we averaged the ELMo embeddings for a given word across the different contexts in which the word appeared. This was necessary because the ERPs for a given word were averaged across contexts. It also put ELMo and Fasttext on more equal footing. However, the single-trial analysis uses the individual instances of each word, with differences in context across these instances, which made it possible to use the corresponding context-specific ELMo word embeddings. When we used the context-specific ELMo embeddings, we found a mean representational similarity (Spearman rho) of 0.0022 ($p < .001$) between the EEG similarity matrix and the ELMo similarity matrix. This was slightly higher than the value of 0.0015 observed when the word embeddings were averaged across contexts, suggesting that the contextual information captured by ELMo correspond with the contextual information used by the human brain.

3.7. Time course analyses

To see the time course of the relationship between the ML-NLP models and the neural responses, we conducted the RSA analysis separate for each time point, using only the scalp distribution information to construct the ERP similarity matrix. The results are shown in Figure 6a, which also includes the lower and upper bounds of the noise ceiling. Just like the correlations for the entire waveform, these single-point correlations were greater for ELMo than for Fasttext. The model-ERP correlations were near zero prior to stimulus onset and then rose rapidly beginning at approximately 100 ms and peaking at approximately 250 ms. Note that small timing errors in our routine for inserting event codes at word onsets may have "smeared out" the time course somewhat and reduced our ability to detect short-latency effects.

The time course results shown in Figure 6a may also be distorted by the baseline correction procedure that is a standard part of our ERP processing pipeline. This correction aims to eliminate low-frequency noise in the data by subtracting the mean voltage during the 500-ms prestimulus baseline period from each epoched EEG waveform. This procedure is usually essential for obtaining reliable ERPs, and it is based on the assumption that the prestimulus period contains only noise (Luck, 2014). However, that assumption is not valid in the present

study, because the baseline period contained ERP activity produced by the preceding words that was potentially predictive of the current word. Consequently, the baseline correction procedure may have artificially reduced the correlation between the representational similarity matrices for the ERPs and the ML-NLP models. Note that although the correlations were small during the baseline period, they were nonetheless statistically significant at many time points (see the horizontal lines at the top of Figure 6a), indicating that the baseline correction procedure did not entirely remove all information about the preceding words.

To assess the influence of the baseline correction procedure, we repeated the time course analyses without baseline correction (relying only on the high-pass filter to attenuate low-frequency voltage offsets). In other words, no baseline correction was applied to the waveforms that were used to create the ERP representational similarity matrices. As shown in Figure 6b, this resulted in lower correlations between the ERP data and the two models than observed in the baseline-corrected data. The noise ceiling was also reduced, consistent with the assumption that baseline correction was helpful in reducing noise. The magnitude of the correlations was reasonably high relative to this reduced noise ceiling. Interestingly, the representational similarity ramped up gradually toward the end of the baseline period when baseline correction was eliminated. This indicates that the brain activity elicited by the preceding words contained information that was predictive of the current word, which could occur many different levels (e.g., coarticulation, syntactic structure, semantic associations).

4. Discussion

This study used representational similarity analysis to compare the representational geometry of word representations in two ML-NLP models (which had been engineered to address practical problems and not to reflect human language processing) with the structure of word representations in the human brain. Although the two ML-NLP models were designed to solve similar problems and were trained with huge sets of natural English text, their representational similarity matrices were only moderately correlated with each other (Spearman rho = 0.49). When averaged across participants, the representational similarity matrix from the neural data exhibited correlations with each of the two models that were only moderately lower (0.30 for the ELMo-ERP correlation; 0.21 for the Fasttext-ERP correlation). Averaging similarity matrices across participants can be problematic (Ashby, 2019; Ashby et al., 1994), but the single-subject correlations were also reasonably strong relative to the noise level of the data.

These results indicate that ML-NLP models of word embeddings—which were designed to solve practical problems such as information retrieval (Nalisnick et al., 2016), sentiment analysis (Shi et al., 2019), machine translation (Zou et al., 2013), and various language understanding tasks (Peters et al., 2018)—end up having a representational geometry that resembles that of the human brain. This is consistent with the hypothesis that any linguistic processing system that can successfully operate on a natural language will end up having a similar underlying representational structure (at least when considered at the abstract level of word-to-word similarities). Comparisons between language embedding spaces for different languages (Conneau et al., 2017; Lample et al., 2017) also support this hypothesis. Such a structure may be inevitable given the semantic relationships between the concepts represented by the words (especially for content words) and the syntactic structure of the language (especially for function words). For example, it is difficult to imagine how an English language processing system could be successful unless it treated “chair” as being more related to “table” than to “velocity” (or “the” as being more related to “a” than to “with”). Acoustic and phonemic

similarities may also play a role. If we assume that the human brain is well suited to processing natural human languages, this suggests that the best NLP systems of the future may have a representational structure that is isomorphic with that of the human linguistic processing system.

Because the two ML-NLP models were only modestly correlated with each other, this created the possibility that each might explain unique variance in the ERP data. Semipartial correlations indicated that only the ELMo model explained significant unique variance in the ERP data. This suggests that the representational structure of ELMo is more similar to that of the human brain than is the representational structure of Fasttext.

However, ELMo has an intrinsic advantage, because the embedding vector produced by ELMo for a given word depends on the local context of the word, whereas Fasttext produces context-independent embedding vectors. Because the ERPs were averaged over all instances of a given word in the two stories to achieve a reasonable signal-to-noise ratio, we also averaged the ELMo embeddings over all of the instances of these words prior to constructing the representational similarity matrix[7]. Nonetheless, the averaged ELMo embeddings did reflect the distribution of meanings in the same stories that were used to elicit the ERP responses, whereas Fasttext did not. In addition, when we conducted a single-trial analysis, we found that ELMo had a stronger representational link to the ERP data when we used the context-specific ELMo embeddings rather than context-independent, averaged ELMo embeddings. However, the single-trial representational similarity values were very small, so this conclusion is tentative.

Although our primary analyses combined the spatial and temporal features of the ERP data to construct the representational similarity matrices, we also tracked the time course of the representational similarity between the neural data and the two models. As can be seen in Figure 6, the time-course of the similarity began ramping up at approximated 100 ms, peaked at approximately 250 ms, and was sustained for several hundred milliseconds before gradually declining. This time course fits with the timing of traditional ERP components that have been used to study lexical and semantic processing. In particular, the N200 family of components is sensitive to auditory word form information (Boudewyn et al., 2015; Connolly & Phillips, 1994; Diaz & Swaab, 2007), and the N400 component is sensitive to word-level associative and semantic information (Kutas & Federmeier, 2011; Swaab et al., 2012). The present effects may reflect, in part, the same neural systems that produce these ERP components, but this is difficult to determine from the present data. It is also worth noting that the early onset of the representational similarity suggests that acoustic or phonemic representations may be associated with the word embeddings of the ML-NLP models.

A recent fMRI study of the cortical distribution of semantic features (Huth et al., 2016) used word embeddings to predict the BOLD response measured while participants listened to stories. To create a data-driven model of semantics, a simple model of word embeddings was constructed that counted the co-occurrence of a given word with every other possible word in a large corpus. Equal weight was given to word pairs separated by up to 15 words to emphasize semantic relationships and minimize syntactic influences on the embeddings. The vector of features for each word was then combined with the sequence of words in the stories to create a set of regressors for predicting the BOLD response. This procedure yielded highly detailed projections of four semantic dimensions (derived from principal component analysis) onto the cortical surface. Although both the present study and this prior study examined the relationship

[7] Given the low signal-to-noise ratio of the single-trial EEG data, it was not realistic to compare the separate ELMo representations for each individual word context to the corresponding single-trial EEG data.

between word embeddings and neural responses to natural language, the goals and analytic methods of the two studies were quite different. The prior study used a simple model of word embeddings that was intended to reflect human semantics, and the set of words was dominated by semantically rich *content* words. By contrast, the present study focused on the 100 most commonly occurring words, which were predominantly *function* words (see Table 1), and it focused on sophisticated models of word embeddings that were developed to solve applied NLP problems rather than to reflect human semantic relationships. In addition, whereas the goal of the prior study was to assess the tiling of semantic features across the cortical surface, the goal of the present study was to quantify the similarity between the representational structure of the ML-NLP models and the human brain. Thus, the present study and the previous study provide complementary information.

From another perspective, the representational similarity between the neural data and the word embedding models can be seen as an index of the isomorphism between the semantic map of the brain and the semantic structure of the embedding space. Frankland & Greene (2020) proposed that the brain could be using a grid-cell like system to represent meanings, resulting in a semantic map in the brain. However, unlike 2-D spaces that can be mapped to a 2-D cortical space, semantic maps have an undefined and high dimensional geometry. It is also very likely that the triangle inequality does not apply in the semantic map. For example, “rodent” and “keyboard” are both semantically similar to “mouse”, but they are very dissimilar with each other. Consequently, it would be difficult for a grid-cell like representation to capture the full extent of the semantic space. RSA does not assume a 2-D representation and can therefore be used to assess the isomorphism between the semantic map of the brain and the complex, large-dimensional semantic structure of the embedding space.

References

Ashby, F. G. (2019). *Statistical analysis of fMRI data* (Second Edition). The MIT Press. http://mitpress.mit.edu/9780262042680

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the Dangers of Averaging Across Subjects When Using Multidimensional Scaling or the Similarity-Choice Model. *Psychological Science*, *5*(3), 144–151. https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Boudewyn, M. A., & Carter, C. S. (2018). I must have missed that: Alpha-band oscillations track attention to spoken language. *Neuropsychologia*, *117*, 148–155. https://doi.org/10.1016/j.neuropsychologia.2018.05.024

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607–624. https://doi.org/10.3758/s13415-015-0340-0

Brown, C. M., Hagoort, P., & ter Keurs, M. (1999). Electrophysiological Signatures of Visual Lexical Processing: Open-and Closed-Class Words. *Journal of Cognitive Neuroscience*, *11*(3), 261–281. https://doi.org/10.1162/089892999563382

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *ArXiv:1312.3005 [Cs]*. http://arxiv.org/abs/1312.3005

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word Translation Without Parallel Data. *ArXiv:1710.04087 [Cs]*. http://arxiv.org/abs/1710.04087

Connolly, J. F., & Phillips, N. A. (1994). Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256–266. https://doi.org/10.1162/jocn.1994.6.3.256

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Diaz, M. T., & Swaab, T. Y. (2007). Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. *Brain Research*, *1146*, 85–100. https://doi.org/10.1016/j.brainres.2006.07.034

Doyle, A. C. (1905). *The Return of Sherlock Holmes*. George Newnes Ltd.

Doyle, A. C. (1992). *The Adventures of Sherlock Holmes*. Wordsworth Editions.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, *20*(1), 116–137.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (Special Volume of the Philological Society)*, *1952–59*, 1–32.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 197–202. https://doi.org/10.3758/BF03204765

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307. https://doi.org/10.1080/01638539809545029

Frankland, S. M., & Greene, J. D. (2020). Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology*, *71*(1), 273–303. https://doi.org/10.1146/annurev-psych-122216-011829

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, *14*(7), e1006327. https://doi.org/10.1371/journal.pcbi.1006327

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434–447. https://doi.org/10.1016/j.neuroimage.2018.02.044

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hill, F., Reichart, R., & Korhonen, A. (2014). SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *ArXiv:1408.3456 [Cs]*. http://arxiv.org/abs/1408.3456

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Kiat, J. E., Hayes, T. R., Henderson, J. M., & Luck, S. J. (under review). Rapid extraction of the spatial distribution of physical saliency and semantic informativeness from natural scenes in the human brain. *Manuscript under Review*.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition*, *11*(5), 539–550. https://doi.org/10.3758/BF03196991

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only. *ArXiv:1711.00043 [Cs]*. http://arxiv.org/abs/1711.00043

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique, Second Edition*. MIT Press.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26, pp. 3111–3119). Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Nalisnick, E., Mitra, B., Craswell, N., & Caruana, R. (2016). Improving Document Ranking with Dual Word Embeddings. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 83–84. https://doi.org/10.1145/2872518.2889361

Neville, H. J., Mills, D. L., & Lawson, D. S. (1992). Fractionating Language: Different Neural Subsystems with Different Sensitive Periods. *Cerebral Cortex*, *2*(3), 244–258. https://doi.org/10.1093/cercor/2.3.244

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep Contextualized Word Representations*. 2227–2237. https://doi.org/10.18653/v1/N18-1202

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Shi, W., Chen, M., Zhou, P., & Chang, K.-W. (2019). Retrofitting Contextualized Word Embeddings with Paraphrases. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1198–1203. https://doi.org/10.18653/v1/D19-1113

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related ERP components. In *The Oxford handbook of event-related potential components.* (pp. 397–439). Oxford University Press.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1393–1398. https://www.aclweb.org/anthology/D13-1141

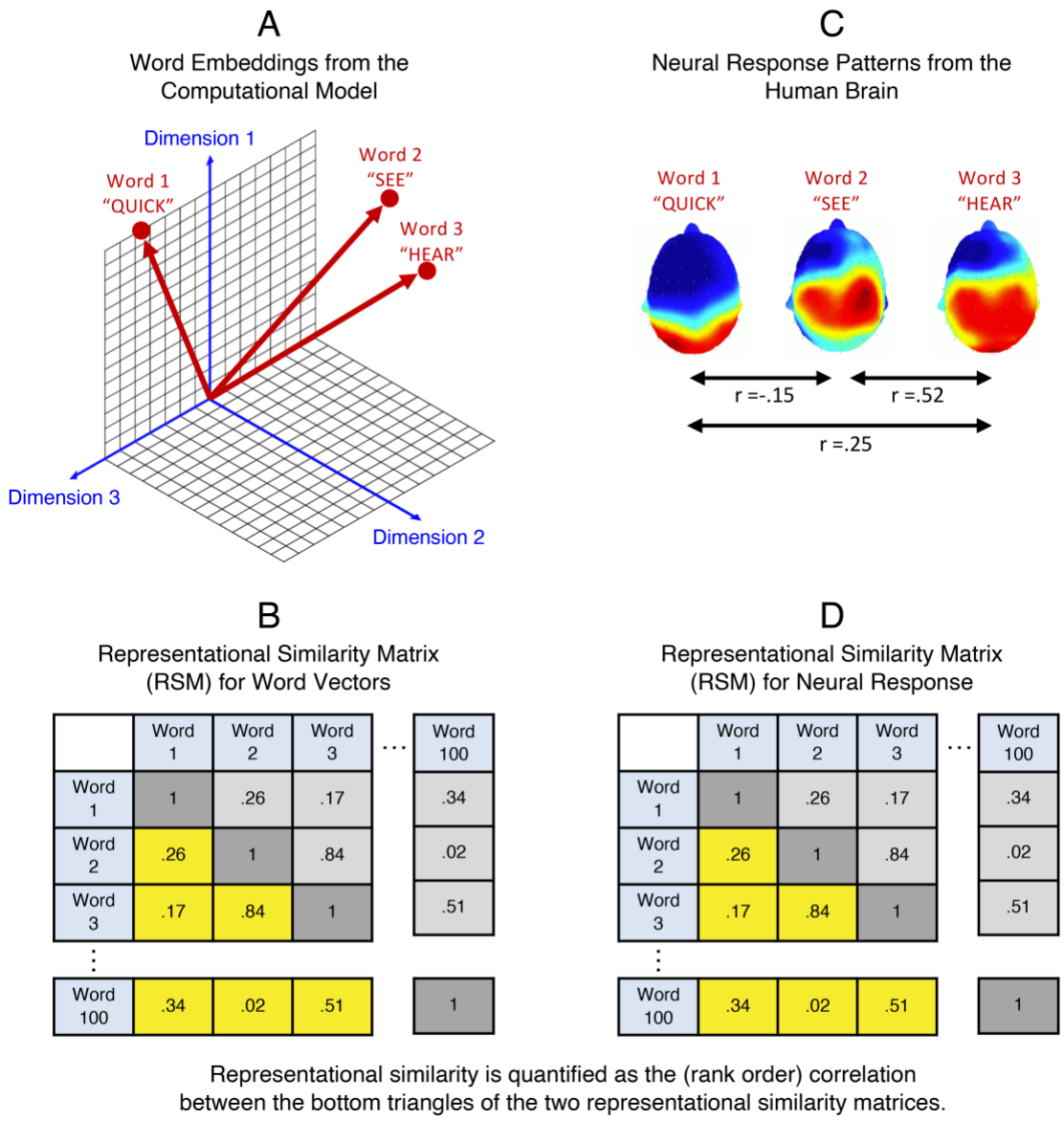

Figure 1. Conceptual overview of representational similarity analysis in a study with 100 different words. (A) **Computational model.** Each word is fed into a computational model, which codes the word as a location in a multidimensional space. Only three dimensions are shown here, but most models have hundreds of dimensions. The similarity between two words can be quantified as the distance between their representations in this space. (B) **Representational similarity matrix from the computational model.** Each cell of the matrix represents the similarity between a given pair of words in the model. (C) **Neural responses.** In this example, the neural response to a word is the distribution of voltage values across scalp electrodes, but different kinds of neural responses can be used (e.g., the pattern of activity across voxels). The similarity between the neural responses to two words can be quantified as the correlation between the corresponding voltage distributions. (D) **Representational similarity matrix from the neural data.** Each cell of the matrix represents the similarity between the neural responses elicited by a given pair of words. The representational similarity between the model and the brain is quantified as the correlation between the matrix for the model and the matrix for the neural data.

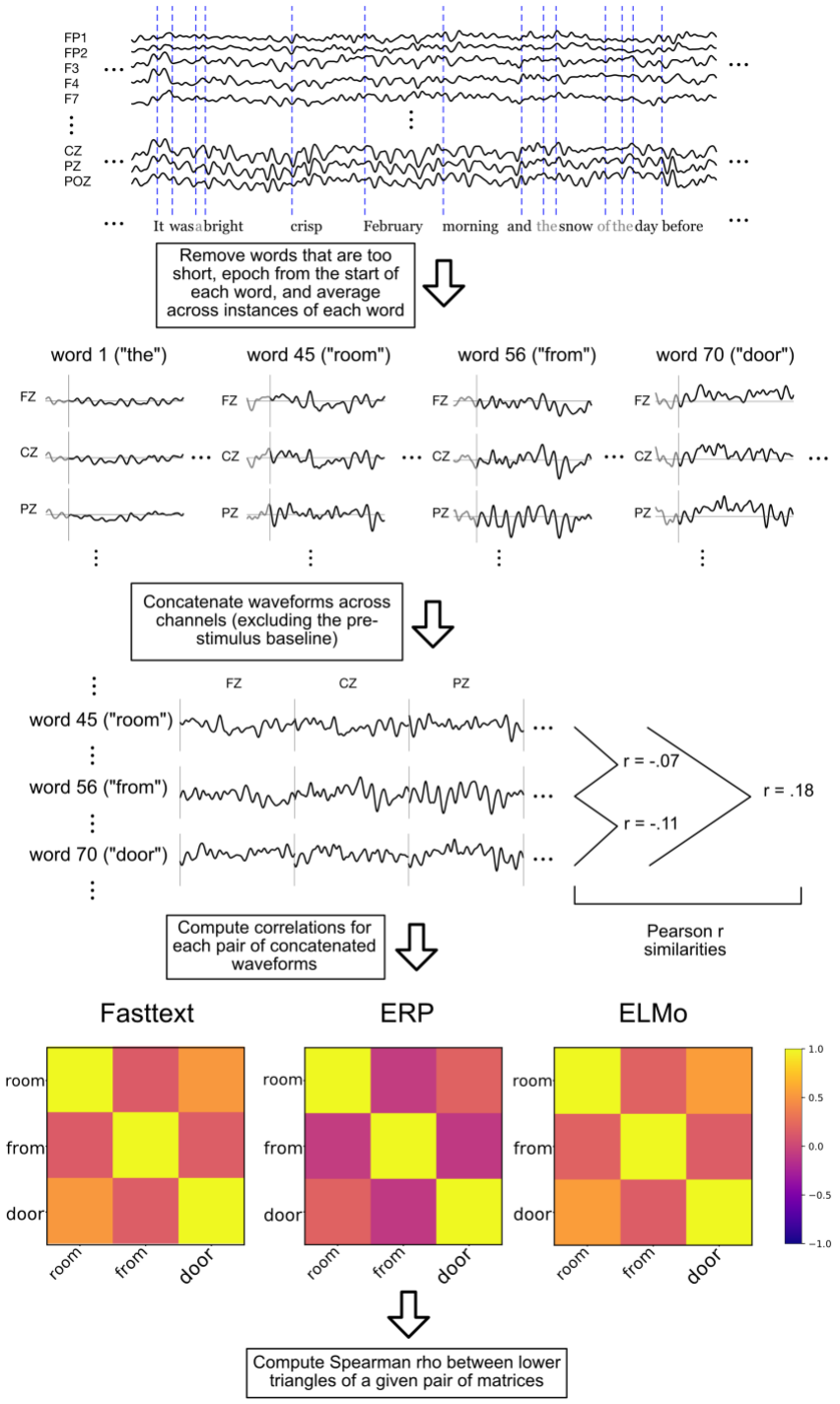

Figure 2. Overview of our procedure for computing representational similarity. For each word, the ERP waveforms at different scalp sites were concatenated together to make a single vector of voltage values. We then computed the correlation between these vectors for each pair of words to create an ERP representational similarity matrix. Finally, we computed the rank-order correlation between this matrix and the representational similarity matrices for the two models to estimate the representational similarity between the ERPs and the models.

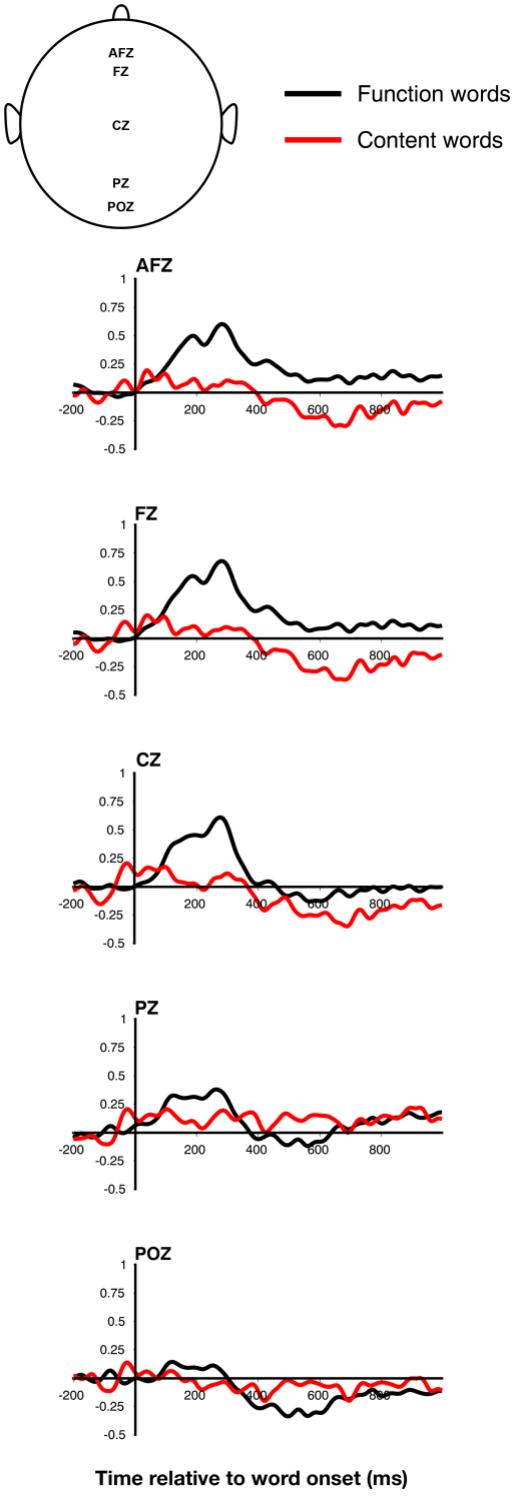

Figure 3. Grand average ERP waveforms, collapsed across all the function words or all the content words, at the midline electrode sites. Time zero is word onset.

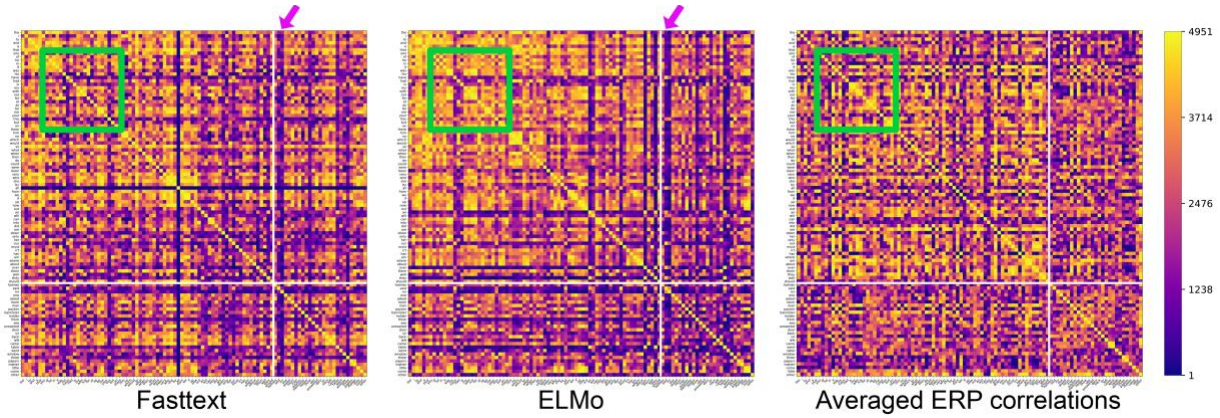

Figure 4. Representational similarity matrices for the Fasttext model, for the ELMo model, and for the ERP data (averaged across participants). The white lines separate the function words (left and top) from the content words (right and bottom). Within each of these categories, the words are ordered in terms of the number of instances within the two stories (with the same ordering from top to bottom as in Table 2). The similarity values have been converted to ranks (with a higher rank meaning greater similarity). That is, the shading within each cell reflects the similarity between the pair of words for that cell, converted into a rank ordering. The purple arrows indicate columns that are obviously similar in the Fasttext and ELMo matrices, and the green squares indicate regions that are visibly similar across all three matrices.

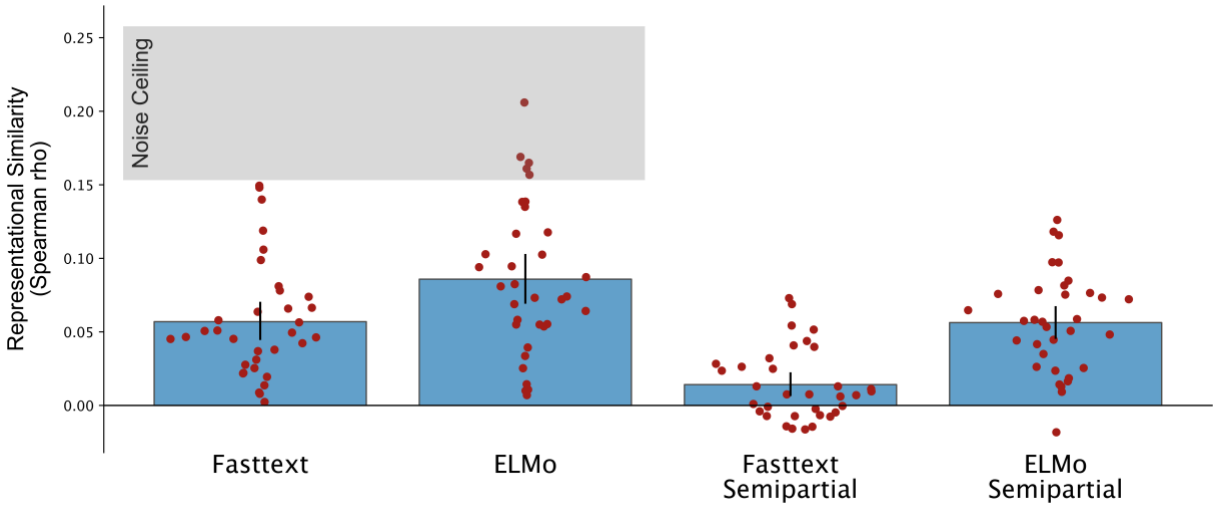

Figure 5. Simple correlations and semipartial correlations between the single-participant ERP similarity matrices and the Fasttext and ELMo similarity matrices. Each dot represents one participant, and the bars represent the mean across participants. The error bars show bootstrapped 95% confidence intervals. The horizontal band across the top shows the upper and lower bounds of the noise ceiling (Nili et al., 2014), which represents the greatest estimated correlation that could be expected given the noise in the data.

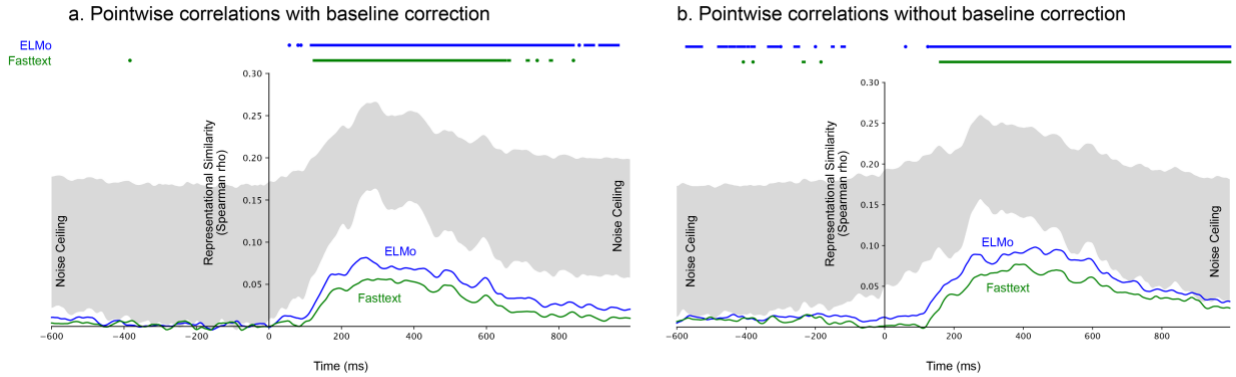

Figure 6. Representational similarity between the ERP data and each of the two models for each time point of the ERP waveform. The representational similarity matrices for the ERP data were computed either with (a) or without (b) baseline correction of the ERP waveforms. Representational similarity was computed separately for each participant, and the mean across participants is shown here. The gray region shows the lower and upper bounds, respectively, of the noise ceiling. The horizontal line segments across the top indicate time periods in which the representational similarity values were significantly greater than zero (after correcting for the false discovery rate).

Supplementary Materials for

Neural Correlates of Word Representation Vectors in Natural Language Processing Models: Evidence from Representational Similarity Analysis of Event-Related Brain Potentials

Taiqi He, Megan A. Boudewyn, John E. Kiat, Kenji Sagae, and Steven J. Luck

Contents

S1. Insertion of Event Codes at Word Onsets

In the original study (Boudewyn & Carter, 2018), each sentence was stored as a separate audio file, and an event code was sent from the stimulus presentation system to the EEG recording system at the onset of each sentence. The original study was designed to examine the ERPs to a subset of words in relation to the attentional state of the participants. Therefore, the stories were periodically interrupted by questions asking the participants about their current attentional state. A total of 54 of these questions appeared across the two stories. Event codes were manually inserted into the data files offline to mark the onsets of content words in the sentences immediately preceding and following these attention questions. Event codes were inserted for 919 words in a total of 164 sentences. Word onset was determined by a combination of auditory inspection and visual inspection of the speech waveform by one of the authors (M.A.B.).

In the present study, we were interested in the ERPs for all the words. Given that there were over 10,000 different word tokens, we used an automated process to insert event codes at the word onsets. Specifically, the audio file for a given sentence was fed into the *Montreal Forced Aligner* (McAuliffe et al., 2017), which is a well-validated tool for determining onset times from recordings of natural speech. The speech signal is converted into a set of acoustic features (mel-frequency cepstral coefficients), and these features are combined with a model pretrained on hand-labeled speech examples to provide estimates of word onsets for new speech inputs (taking into account coarticulation).

We initially validated the results of this process by inspecting the segmentations of several sentences manually, and we found no obvious errors. We then compared the outputs with the hand-coded word onsets. On average, the error was -8.88 ms, meaning that the computed onsets were biased to be slightly earlier than the manually determined onsets. To quantify the variability of the automated onset estimates relative to the manually determined onsets, we took the absolute value of the error for each word token and computed the mean across tokens. The mean absolute error was 19.64 ms. Figure S1 shows the distribution of errors.

We wrote custom Python code to take these word onset estimates and insert event codes into the EEG data files at the appropriate times, anchored by the start time of each audio file. We then excluded words that were shorter than 100 ms.

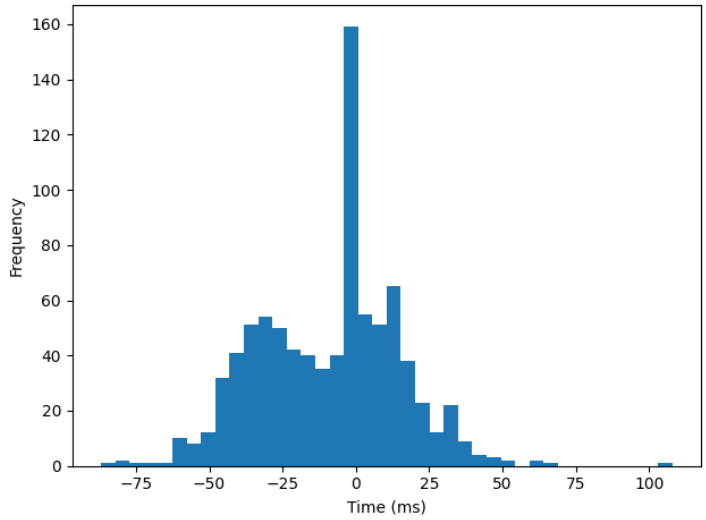

Figure S1: Errors of automatically generated word onsets relative to manually determined word onsets. Negative values indicate that the computed onset was earlier than the manually determined onset..

S2. Effects of Averaging Similarity Matrices Across Participants

Although it is possible to reduce noise in representational similarity matrices (RSMs) by averaging the single-participant RMSs together into a grand average RSM, this approach can lead to invalid conclusions. The mathematical issues are somewhat complex and are described in detail by (Ashby et al., 1994). Here, we provide concrete examples designed to provide an intuitively appreciation of how averaging RSMs can potentially lead to erroneous conclusions.

Below, we provide RSMs for two example participants along with the average of the two RSMs and the RSM from a model. The correlation between Subject 1's RSM and the model RSM is +.61, whereas the correlation between Subject 2's RSM and the model is -.61. Taking the average of those single-participant correlations, one would correctly conclude that there is currently no evidence for a significant positive or negative relationship between the "average" participant's RSM and the model.

However, if one were to instead average across the subject RSMs to create an average RSM, the correlation between that average RSM and the model RSM would be r = +.61, potentially leading to the erroneous conclusion that the subject RSMs are positively related to the model RSM "on average".

Subject 1 RSM

1	.25	.25
.25	1	.3
.25	.3	1

Subject 2 RSM

1	.4	.4
.4	1	.3
.4	.3	1

Average of Single-Participant RSMs

1	0.325	0.325
0.325	**1**	0.3
0.325	0.3	1

Model RSM

1	0.5	0.2
0.5	1	0.15
0.2	0.15	1

S3. References for Supplementary Materials

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the Dangers of Averaging Across Subjects When Using Multidimensional Scaling or the Similarity-Choice Model. *Psychological Science*, *5*(3), 144–151. https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

Boudewyn, M. A., & Carter, C. S. (2018). I must have missed that: Alpha-band oscillations track attention to spoken language. *Neuropsychologia*, *117*, 148–155. https://doi.org/10.1016/j.neuropsychologia.2018.05.024

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386