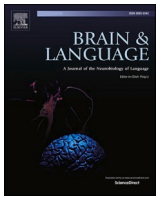
Brain & Language 250 (2024) 105389

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Sentence-level embeddings reveal dissociable word- and sentence-level cortical representation across coarse- and fine-grained levels of meaning

Scott L. Fairhall *

Center for Mind/Brain Sciences (CIMeC), University of Trento, Italy

ARTICLE INFO

Keywords:
Semantics
Concepts
fMRI
Representational similarity analysis
Large language models

ABSTRACT

In this large-sample (N = 64) fMRI study, sentence embeddings (text-embedding-ada-002, *OpenAI*) and representational similarity analysis were used to contrast sentence-level and word-level semantic representation. Overall, sentence-level information resulted in a 20–25 % increase in the model's ability to captures neural representation when compared to word-level only information (word-order scrambled embeddings). This increase was relatively undifferentiated across the cortex. However, when coarse-grained (across thematic category) and fine-grained (within thematic category) combinatorial meaning were separately assessed, word- and sentence-level representations were seen to strongly dissociate across the cortex and to do so differently as a function of grain. Coarse-grained sentence-level representations were evident in occipitotemporal, ventral temporal and medial prefrontal cortex, while fine-grained differences were seen in lateral prefrontal and parietal cortex, middle temporal gyrus, the precuneus, and medial prefrontal cortex. This result indicates dissociable cortical substrates underly single concept versus combinatorial meaning and that different cortical regions specialise for fine- and coarse-grained meaning.

1. Introduction

While the cortical representation of the semantic content of single words has been extensively studied in past decades, it is the capacity to combine these single concepts in increasingly complex combinations that underlies many of the uniquely human aspects of knowledge and thought. One way to probe the cortical representation of such combinatorial meaning is by assessing sensitivity to meaning at the level of the sentence. A promising avenue to study sentence-level semantic sensitivity is offered by recent advances in computational language models. It is the ability of these models to capture meaning across the whole sentence that allows both the formulation of combinatorial representational spaces and the consideration of whether word- and sentence-level meaning are represented differentially across the cortex. In the current work we use OpenAI's sentence-level language models to address these questions and additionally utilise the properties of our stimulus set to assess whether our brain encodes fine-grained differences in sentence meaning between similar sentences and course-grain differences between highly dissimilar sentences in similar or different ways.

Sensitivity to semantic content at the level of the single word has been investigated in terms of local regional selectivity for semantic classes. These localised cortical increases in fMRI response have been observed for classes such as tool-, person- or place-related concepts (Chao, Haxby, & Martin, 1999; Fairhall & Caramazza, 2013a, 2013b; Fairhall, 2020; Fairhall, Anzellotti, Ubaldi, & Caramazza, 2014; Noppeney, Price, Penny, & Friston, 2006; for a review, see Bi, Wang, & Caramazza, 2016), as well as specific semantic features (Fernandino et al., 2015; Liuzzi, Aglinskas, & Fairhall, 2020). At the same time, Multivariate Pattern Analysis (MVPA) has identified more subtle sensitivity to semantic content in the distributed pattern of activation across voxels (Bruffaerts et al., 2013; Devereux, Clarke, Marouchos, & Tyler, 2013; Fairhall & Caramazza, 2013a; Leonardelli & Fairhall, 2022; Liuzzi et al., 2015; Simanova, Hagoort, Oostenveld, & Van Gerven, 2014). MVPA representation can be further extended with Representational Similarity Analysis (RSA), to assess where the distance between neural patterns produced by specific words can be seen to align with the semantic distance between those words. This extension allows the researcher to know not only that there is some form of information present about the different classes of words but that this conforms to a particular property (in this case, semantic meaning). Such neuro-conceptual similarity has been quantified using word similarity derived from co-occurrence with other words (Fu et al., 2023), based on

* Address: Center for Mind/Brain Sciences, University of Trento, Corso Bettini 31, I-38068 Rovereto, Italy.
E-mail address: scott.fairhall@unitn.it.

https://doi.org/10.1016/j.bandl.2024.105389
Received 31 May 2023; Received in revised form 9 January 2024; Accepted 26 January 2024
Available online 2 February 2024

hierarchically defined, Wordnet derived, distances (Fairhall & Caramazza, 2013a, 2013b) or based upon word embeddings (Anderson et al., 2019; Fu et al., 2023; Liuzzi et al., 2020). This latter class consist of numerical vectors derived from linguistic neural networks that capture meaning through the localised context within which words appear in text (e.g. *word2vec*; Mikolov, Chen, Corrado, & Dean, 2013), is the initial input into most extant large language models, and provides a powerful tool for mapping meaning in the brain. Collectively, what emerges from these studies of category-general cortical sensitivity to word meaning is a distributed network of left hemisphere biased regions centred on the precuneus and medial prefrontal cortex (mPFC), the angular gyri (AG), ventral temporal cortex (VTC), posterior middle temporal gyri (pMTG), the inferior frontal gyri (IFG) and the left lateral orbitofrontal cortex (latOFC).

Similar results have been observed when participants are presented with sentence stimuli. Sentences describing particular thematic categories, such as those referring to people, places, object or animals, show pronounced domain-selective responses for these classes (Rabini, Ubaldi, & Fairhall, 2021, 2023; Ubaldi, Rabin, & Fairhall, 2022). When single word embeddings are averaged across a sentence (treating a sentence merely as the sum of its individual words) to create representational spaces, RSA results also closely converge with single word studies (Acunzo, Low, & Fairhall, 2022; see also Pereira et al., 2018). Recent advances in artificial intelligence (AI) models of sentence-level meaning have opened investigation of the neural representation of the higher order meaning conveyed by the structured integration of words into the overall meaning conveyed by the sentence. Models of sentence meaning have been seen to capture neural representations of meaning generally across the language network and have been used to investigate the architectural properties, training goals and neural-network layers that lead to the most neural-like representations, with varying results (Caucheteux & King, 2022; Schrimpf et al., 2021; Sun, Wang, Zhang, & Zong, 2021). However, it is important to note that neither word- or sentence-level approach in isolation can distinguish the cortical response to the individual constituent words from the combinatorial meaning of sentences.

To identify the neural processes underlying the higher-order meaning conveyed by the structured integration of words into sentence-level meaning, it is necessary to compare word-level to sentence-level representations. Region-of-interest analysis comparing deep neural network representations of sentence meaning to the averaged meaning of single words uncovered significantly greater sentence-level representations in the left temporal cortex and bilateral inferior temporal gyrus in compositional meaning (Anderson et al., 2021). In addition, recent work using sentence level representations derived from a deep neural network trained to identify the topic of the sentence show an enhanced ability to capture neural representation compared to averaged single-word embeddings found greater responses additionally in the precuneus and medial prefrontal cortex (Acunzo et al., 2022). However, in this latter case it is not possible to fully distinguish whether this arises due to the contribution of training goals (topic differentiation), the relationship between the sentence stimuli used in the fMRI study and these training goals (which may have led to bias) or from the sentence-level processing capacity conveyed by the deep neural network's architecture.

Sentence-level embeddings offer another promising way to address how semantic representation changes from single words to sentences. Like their word-embedding counterparts, these sentence level embeddings represent the meaning, now of the entire sentence including elements of the structure that it conveys. In the present work, we use OpenAIs text-embeddings-ada-002 to estimate the relatedness of 240 sentences drawn from four broad thematic categories of sentences. We use this in conjunction with representational similarity analysis (RSA) and fMRI to identify maximal convergence between sentence and neural relatedness across the brain. The contribution of sentence-level meaning is isolated by comparison to a model where word order within the sentences have been scrambled. We further informed this analysis by utilising the categorical structure or our stimulus set to examine both coarse-grained (across thematic category) and fine-grained (within thematic category) sentence representation.

2. Methods

2.1. Participants

Data are taken from 64 right-handed participants (26 male; mean age, 24.9) who underwent functional magnetic resonance imaging while reading general-knowledge questions presented in Italian. Data were derived from two studies. The first (N = 24) was a study of transient blocks in semantic access and participants read questions and were asked whether the target piece of knowledge was accessible, thought to be known but presently inaccessible, or was unknown (Ubaldi et al., 2022). In the second study (N = 40), an investigation of individual differences, participants read questions and reported only whether the target piece of knowledge was accessible or not (Rabini et al., 2023). In both cases, the task is orthogonal to the present study, where all sentences are analysed irrespective of response. All participants gave informed consent and procedures were approved by the Ethics Committee at the University of Trento and were conducted in line with the declaration of Helsinki (1964, amended in 2013).

2.2. Stimuli

Stimuli were composed of 240 general-knowledge questions written in Italian. The questions were equally divided into four thematic categories that described: people (e.g., "Which philosopher uttered the phrase "I know that I don't know?"); places, (e.g., "In which Spanish city is the Alhambra complex located?"); objects, (e.g., "What is the name of the brick that supports the weight of an arch?") or a final 'scholastic' knowledge domain that was designed to capture general knowledge unrelated to direct experience with the environment (e.g., "What is the name of the transition of matter from the solid to the gaseous state?"). Stimuli were similar but not identical between the two studies (94 % were either the same or differed minimally in wording). Sentences were on average 10 words in length (study 1: mean: 9.95, range 6–15; study 2: mean: 10.0, range 8–12) and were matched across knowledge domain by number of words and number of letters. In both study 1 and 2, sentences were only presented once per participant. The full list of stimuli is available for study 1 here: https://figshare.com/s/2f4be0ba0278ea79a7d5; and for study 2, here: https://figshare.com/s/de4c9df3a39fc16fe0d3.

2.3. Procedure

Stimuli were presented using Matlab (https://www.mathworks.com) and Psychtoolbox Version 3 (https://www.psychtoolbox.org). In both studies, 15 questions from each of the four knowledge domains were presented per run in a pseudorandomised event-related design. In study 1, sentence stimuli were presented for 3 s followed by a 3 s fixation cross. In study 2, questions were presented one word at a time for 250 msec each, in black with the left-of-centre letter printed in red (≈17 % left-of-centre), in order to facilitate fixation. For this study, presentation time was 2–3 s (depending on sentence length) and was followed by a fixation cross for the remaining duration of the six-second trial. In both studies, participants responded whether the targeted piece of knowledge was fully accessible in that moment or not. In study 1, if the question's answer was not accessible, participants were subsequently asked their confidence that they possessed the knowledge. Thus, the sentences analysed in the present study consisted of trials where the targeted answer was available to the participant in the moment ('accessible' response) or situations where the targeted piece of knowledge was either unknown, transiently inaccessible, or only partially accessible.

2.4. MRI scanning parameters

Functional and structural data were collected at the Center for Mind/Brain Sciences (CIMeC), University of Trento, with a Prisma 3 T scanner (Siemens), using a 64-channel head coil. Participants lay in the scanner and viewed the visual stimuli through a mirror system connected to a 42 in., MR-compatible LCD monitor (NordicNeuroLab) positioned at the back of the magnet bore. Functional images were acquired over four runs using echoplanar T2*-weighted scans. Run duration was on average 10.5 min in study 1 (depending on responses) and 7 min in study 2. Acquisition parameters were as follows: repetition time (TR), 2 s; echo time (TE), 28 ms; a flip angle, 75°; field of view (FoV), 100 mm; matrix size, 100 × 100. Each volume consisted of 78 axial slices (which covered the whole brain) with a thickness of 2 mm, anterior commissure/posterior commissure aligned.

High-resolution (1 × 1 × 1 mm) T1-weighted MPRAGE sequences were also collected (sagittal slice orientation; centric phase encoding; image matrix, 288 × 288; FoV, 288 mm; 208 slices with 1 mm thickness; TR, 2290 s; TE, 2.74 ms; inversion time, 950 ms; flip angle, 12°).

2.5. fMRI data analysis

Data were analysed and preprocessed with SPM12 (https://www.fil.ion.ucl.ac.uk/spm/). The first four volumes of each run were dummy scans. All images were slice time corrected, realigned to correct for head movement, normalized to MNI space, and smoothed using a 6 mm FWHM isotropic kernel. The data were then temporally high-pass filtered using custom code and a FIR filter of order 80, and cut-off = 0.0156 Hz (64 s). To estimate the fMR response for each trial was attained by averaging the 3 EPI volumes between 6 and 10 s post stimulus onset (in 0.5 % only two were used as the third scan was unavailable). Analyses were performed within an a priori defined grey matter mask.

2.6. Language models

Sentence embeddings were extracted from OpenAI's second generation text-embedding-ada-002 model (Neelakantan et al., 2022; https://api.openai.com/) for each Italian sentence. Sentence-level embeddings are trained to learn a low dimensional representation that efficiently and numerically captures semantic meaning. The ADA-002 embeddings are derived from an unsupervised contrastive learning approach (Neelakantan et al., 2022), where the transformer encoder learns to produce similar embedding for sentences that occur next to one another in a passage of text and dissimilar embeddings for sentences that do not. The embeddings were used to determine sentence similarity via correlation between the 1536 element embedding vectors of each sentence. The efficacy of this approach as applied to our stimulus sets can be seen in the tight conformation of embedding-similarity based clustering and experimenter-defined categories shown in Fig. 1A.

Six different sentence-embedding derived template models were employed to construct representational dissimilarity matrices (RDMs;

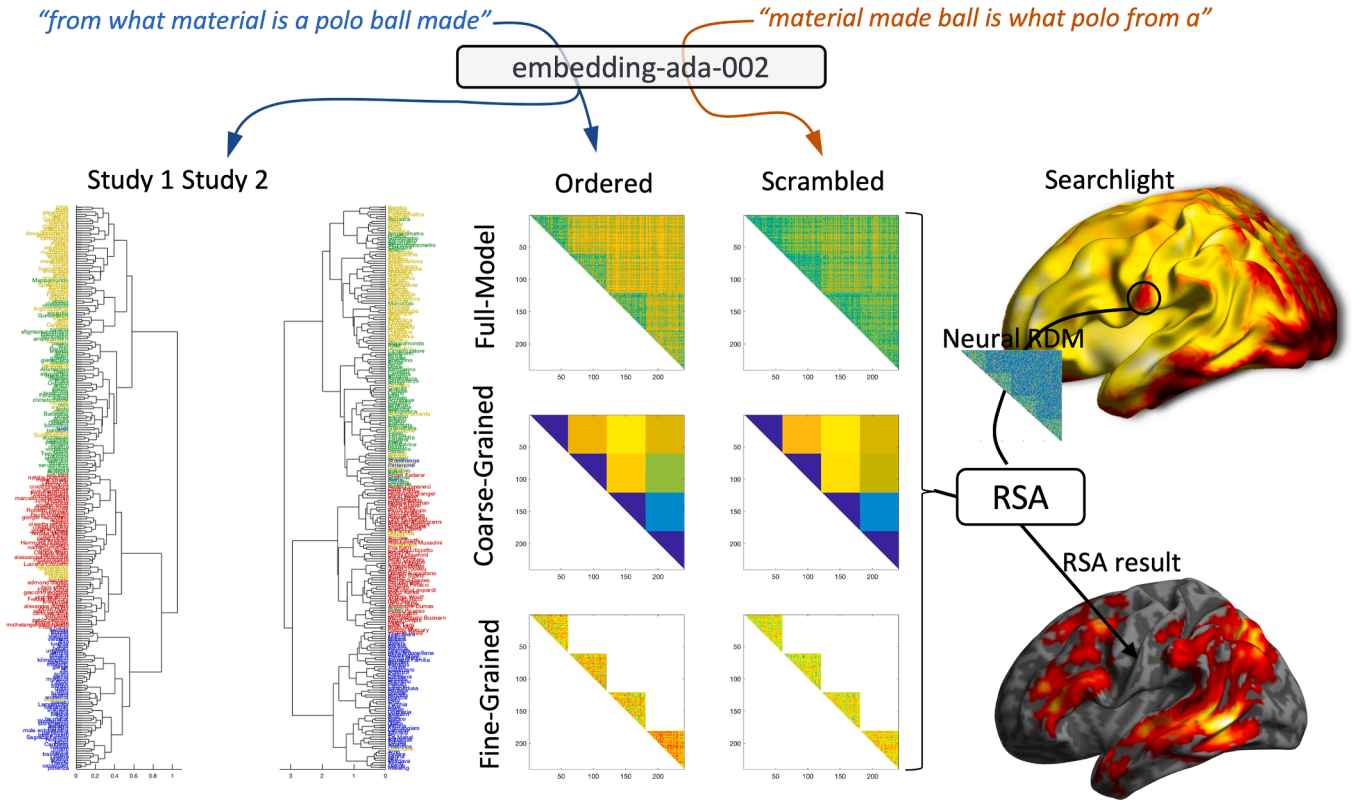


Fig. 1. Semantic representational spaces were created for the 240 sentences using the ada-002 text embeddings. **A.** Stimuli and Semantic Space. Separate Hierarchical clusterings of the embeddings of stimuli from study 1 and study 2 with the sentence topic (indicated by the question's answer) colour-coded by the original experimenter-defined thematic category (red: person; blue: place; green: object; yellow: scholastic). **B.** Template RDMs. To isolate the combinatorial meaning contained within sentence structure from the meaning conveyed by the constituent words in isolation, embeddings were attained for each sentence either in their original form ('ordered') or a word-order shuffled ('scrambled') form to create separate models. Models were further separated in the full-model (all-sentences), coarse-grained model (where each sentence's embedding was replaced by the average of that domain) and fine-grained model (where RSA was performed separately within each knowledge domain and the results averaged). This resulted in six RDMs (in addition to an 'uniformed' binary category RDM, see text). **C.** Searchlight RSA. Separately for each subject and each template RDM, searchlight RSA was performed by correlating the neural RDMs of the 240 sentences extracted from a 4-voxel radius sphere with the template RDM. This process was repeated iteratively with a searchlight sphere centred at each voxel with the resulting template-neural RDMs correlation summarised at the central voxel for that sphere.

see Fig. 1B). To isolate the effects of sentence-level meaning from that of the constituent words alone, models were derived from a) the sentences presented in their original order and b) from sentences where the word order had been randomly scrambled. These ordered and scrambled-order models were used with a full model which included the distance calculated for each sentence pair. Additionally, we utilised the categorical structure of our stimulus set to examine both coarse-grained and fine-grained sentence representations. For the coarse-level model, the embedding for each sentence within a category was replace by the average embedding of that category. This permitted the isolation of the contribution of the broad distance between sentence knowledge-domains by removing fine-grained differences between similar sentences. We additionally included an 'uninformed' coarse-grained model (not shown) where categories were simply assigned a distance of same ('0') or different ('1'). For the fine-grained model, to isolate brain regions most sensitive to subtle differences between more like sentences, RDMs were formed separately for each category. Collectively, this resulted in seven separate template RDMs.

Word-order scrambled model RDMs were seen as the tightest control for the ordered model RDMs, as the sentences are processed by the same model in the same way with the only exception being that word order has been rendered uninformative through scrambling. However, it may be the case that scrambling introduces unanticipated effects that make this a poor model. To control for this possibility, the RDMs derived from the scrambled-order model were compared to the averaged embeddings of the single words contained in each sentence using both the same embeddings (ada-002) and GloVe embeddings (Pennington, Socher, & Manning, 2014; see Acunzo et al., 2022, for training details). In both cases, the word-order scrambled model was superior to the averaged-embeddings models analysis (see supplementary Fig. S1) confirming that the former provides an appropriate control condition for this study.

2.6.1. Representational similarity analysis

RSA was performed by comparing neural representational dissimilarity matrices (RDMs; see Fig. 1C) to language-model derived template RDMs utilising CoSMoMVPA (https://www.cosmomvpa.org/; Oosterhof, Connolly, & Haxby, 2016). Neural RDMs were extracted via searchlight (Kriegeskorte, Goebel, & Bandettini, 2006), where the local pattern of voxel activation was extracted from a 4-voxel radius sphere for each of the 240 sentences. The (dis)similarity between the neural representation of the 240 sentences was then determined by Pearson correlation (1-r), which formed the neural RDM at that location. The concordance between this neural RDM and template RDMs was calculated via Pearson's correlation and the resulting value was recorded at the voxel at the centre of this searchlight. This process was repeated for a sphere centred at every location within the brain volume. This process was repeated for each template RDM resulting in seven searchlight maps for each subject. The brainmaps reflected the results of the ordered and scrambled versions of the full, course-grained and fine-grained models as well as the uninformed binary category model. For analysis of fine-grained differences, the RSA was performed separately within each category and the results of the four resulting analyses averaged.

To perform inferential group level analysis, searchlight maps for each participant were taken in separate random effects group-level analyses to assess word-ordered and scrambled ordered variants of the full, coarse-grained and fine-grained models and for the comparison of the informed and uninformed coarse-grained models. By performing these three separate analyses for the full, coarse-grained and fine-grained models, the degree of within-RDM averaging was balanced at each contrasted level (which may potentially affect the amount of noise present in the RDM).

2.6.2. Region of Interest (ROI) Analysis

ROI analysis was performed to assess differentiated regional contributions to sentence-level meaning across the cortex. Orthogonal ROIs were defined for the full model using the average of the ordered- and scrambled-sentence RSA result. While direct comparisons to baseline would not be valid, the resulting ROI definition is unbiased in terms of differences between the two models and differences across regions. This is because the defining contrast contains no Information about the difference between the two models and therefore cannot bias voxel selection towards either model (for discussion, see Friston, Rotshtein, Geng, Sterzer, & Henson, 2006). To form the ROIs, firstly the 16 most significant peaks in the group-level analysis of the average of the ordered-sentence RSA and the scrambled-sentence RSA were identified. Then the ROIs were defined as the conjunction between a 5 mm spherical ROI centred at each location and voxels that showed significant effect ($p < .001$) in this contrast.

3. Results

3.1. Full model

Sentence-level neuroconceptual similarity was extensive, with a single significant cluster encompassing much of the cortex (Fig. 2A). Within this cluster, the correspondence between modelled sentence similarity and neural similarity spaces was more pronounced in the left hemisphere. Peaks can be broadly grouped into two subtypes. The first includes retrosplenial complex (RSC), posterior parahippocampal gyrus (pPG) and in an anterior section of transverse occipital sulcus abutting the posterior angular gyrus (TOS/AG). These regions have previously shown to have a univariate preference for place-referent sentences (Ubaldi et al., 2022, Rabini et al., 2023) or pictures of places (Epstein & Kanwisher, 1998), to respond when participants have to navigate spatial location within memory (Epstein, Parker, & Feiler, 2007), or when participants retrieve knowledge about specific buildings (Fairhall, Anzellotti, Ubaldi, & Caramazza, 2014) or types of buildings (Fairhall & Caramazza, 2013a, 2013b) and show increased activity when individual have to recall the geographic provenance of specific items such as famous food dishes or people (Fairhall, 2020). Conversely, these regions are not commonly reported in studies investigating the multivariate representation of word meaning (Acunzo et al., 2022; Leonardelli & Fairhall, 2022; Liuzzi et al., 2020). It is of note that the conventional designation of this place-selective region as the 'retrosplenial complex' is somewhat of a misnomer as this region does not encompass the retrosplenial cortex and is rather located in the medial parietal occipital sulcus (Silson, Steel, & Baker, 2016), a region not commonly associated with general semantic processing. Likewise, the region of TOS identified in this study is approximately 1.5 cm posterior to AG regions previously shown to be sensitive to semantic content (Acunzo et al., 2022; Leonardelli & Fairhall, 2022; Liuzzi et al., 2020) and the bilateral pPG seen in the present study are approximately 1 cm posterior to the left lateralised section of the VTC that exhibits a robust multivariate representation of semantic meaning. The second subtype of peaks includes the pMTG, IFG extending into the middle frontal gyrus (MFG), anterior superior temporal sulcus (aSTS), lateral OFC, supramarginal gyrus (SMG), which have previously been seen to exhibit strong multivariate representations of word/sentence meaning (Acunzo et al., 2022; Leonardelli & Fairhall, 2022; Liuzzi et al., 2020). This pattern was largely preserved when considering representational models based on word-order scrambled sentences (Fig. 2B), indicating that much of the capacity of sentence embeddings to capture neural representational spaces is contingent on the presence of the specific words themselves, rather than how they are constructed into a meaningful sentence.

Examination of the difference between the word-ordered and word-order scrambled models allows differentiation of the meaning endowed by the higher-order structure of sentences from the collective response of the composite words. The ordered model significantly outperformed the scrambled model across the brain, indicating distributed contributions of sentence level information to cortical representation (Fig. 2C). Notably, peaks showing the greatest difference between ordered and scrambled models correspond closely to those regions showing the

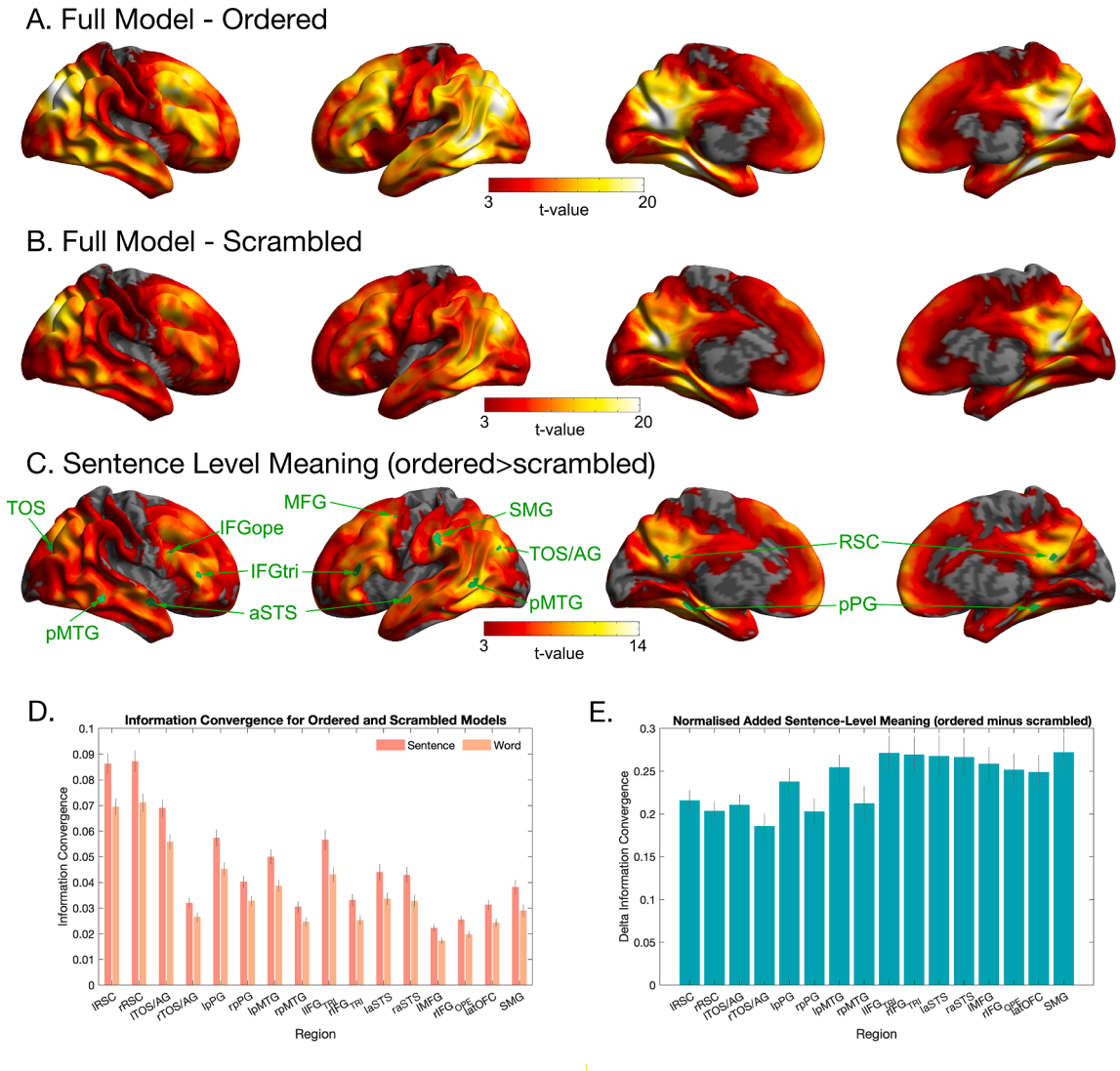


Fig. 2. Results of RSA using the full model that was derived from each sentence's individual embedding for all sentences. **A.** Model created from sentences with words in canconical order. **B.** Model created by word-order scrambled sentences. **C.** Differences between ordered and scrambled brain maps. **D.** Infromational convergence for ordered and scrambled models within ROIs. **E.** Difference between ordered and scrambled models normalised by average regional infromational content (see text). All brain maps are shown with an initial voxel theshold of p <.001, FWE-corrected for multiple comparisons at the cluster level (p <.05). *Abbreviations: IFGope: par operculum; IFGtri: par triangularis.* See Table 1 for peak locations and significance.

maximal overall effect for ordered and scrambled models (compare peaks in Fig. 2A-C), suggesting a relative lack of cortical differentiation.

To assess whether subtle regional variations in sentence-level representation were present, unbiased ROIs were defined through the contrast of the average of full and scrambled model (see methods). An initial ANOVA considering the difference in information captured by the ordered and word-scrambled models indicated strong regional variations ($F_{(14,63)} = 41.36$, $p < .0001$). However, this absolute difference may reflect a difference in the underlying information present in each region (c.f. Fig. 2D; Henson, 2006). To assess whether sentence-level information confers the same proportional increase in neural information capture, we normalized each region by the average amount of information present for the ordered and scrambled models (Fig. 2E). Regional differences were seen to persist ($F_{(14,63)} = 3.88$, $p < .0001$), with the most notable difference being between the smaller increase in putative place-selective regions (PPG, RSC, TOS/AG; mean increase 20.9 %), compared to non-selective regions more closely associated with semantic processing (IFG, aSTG, pMTG, SMG, latOFC; mean increase 25.7 %, $t(63) = 5.36$, $p < .00001$). However, no differences were seen within this non-selective set of regions ($f_{(9,63)} = 1.15$, $p = .33$). Thus, sentence-level meaning captures more information across the cortex but, while this was less pronounced in regions that have been associated with place-information selectivity, the effect was largely undifferentiated in regions associated with general semantic processing.

3.1.1. *Coarse-grained semantic representation*

Our sentence stimuli were drawn from four thematic categories (see methods). This allowed us to examine coarse-grained differences between thematically distinct sentences, in contradistinction to the more fine-grained differences that may distinguish more similar sentences. To accomplish this, RSA was performed using a model where the embedding of each sentence within a category was replaced by the average embedding of that category (Fig. 3A). This model was then compared to an uninformed model that assumed the categories were equidistant from one another (Fig. 3B). Notably, the loci of peak neuroconceptual similarity were consistent with the full model (c.f. Fig. 2A/B), indicating the relative importance of these coarse-grained semantic differences in capturing neural representational spaces.

To isolate the regions where the semantic distance between categories is consistent with that of the sentence embeddings, the difference

Table 1
Peak significance for word-ordered and word-order scrambled models, and the difference between them. Peak locations in MNI coordinates are shown for these contrasts and for the localiser contrast used to define the ROIs. Peaks for each contrast are taken from a single cortex-spanning cluster (extent in voxels: 175992, ordered; 163448, scrambled; 138861, difference).

	Peak T			Peak Location			
Region	**Ordered**	**Scrambled**	**Difference**	**Ordered**	**Scrambled**	**Difference**	***Localizer***
lRSC	27.53	22.25	13.52	−10 −60 20	−12 −60 20	−6 −60 20	*−10 −60 20*
rRSC	27.44	22.38	12.69	12 −54 20	14 −54 20	10 −54 20	*12 −54 20*
lTOS/AG	25.29	20.49	11.64	−34 −78 28	−32 −78 28	−36 −78 30	*−34 −78 28*
rTOS/AG	24.21	20.16	10.3	28 −72 36	28 −72 36	28 −62 44	*28 −78 28*
lPPG	22.87	18.35	12.54	−30 −40 −18	−30 −40 −18	−34 −26–22	*−30 −40 −18*
rPPG	22.46	18.04	10.3	30 −38 −16	30 −38 −16	32 −40 −18	*30 −38 −16*
lpMTG	21.76	16.82	12.2	−52 −58 2	−52 −58 2	−50 −58 2	*−52 −58 2*
rIFG_OPE	20.09	15.57	11.28	44 14 30	42 14 30	46 14 28	*42 14 30*
lMFG	19.18	14.83	10.51	−32 4 50	−32 4 50	−34 6 48	*−32 4 50*
latOFC	18.87	14.72	9.99	−36 32 −12	−36 32 −12	−36 32 −12	*−36 32 −12*
lIFG_TRI	18.18	13.84	10.51	−46 28 12	−46 28 12	−38 30 14	*−46 28 12*
raSTS	17.32	13.26	9.77	56 −4 −14	56 −4 −14	56 −4 −14	*56 −4 −14*
rIFG_TRI	17.11	131.14	9.98	46 36 12	46 36 14	48 30 8	*46 36 12*
rpMTG	17.11	13.85	–	56 −42 −10	58 −42 −10	–	*56 −42 −10*
laSTS	17.11	13.12	9.6	−56 −8 −12	−56 −8 −12	−56 −8 −12	*−56 −8 −12*
SMG	16.85	12.83	9.83	−58 −30 36	−58 −30 36	−60 −30 34	*−58 −30 36*
mPFC	–	–	9.81	–	–	6 58 16	–

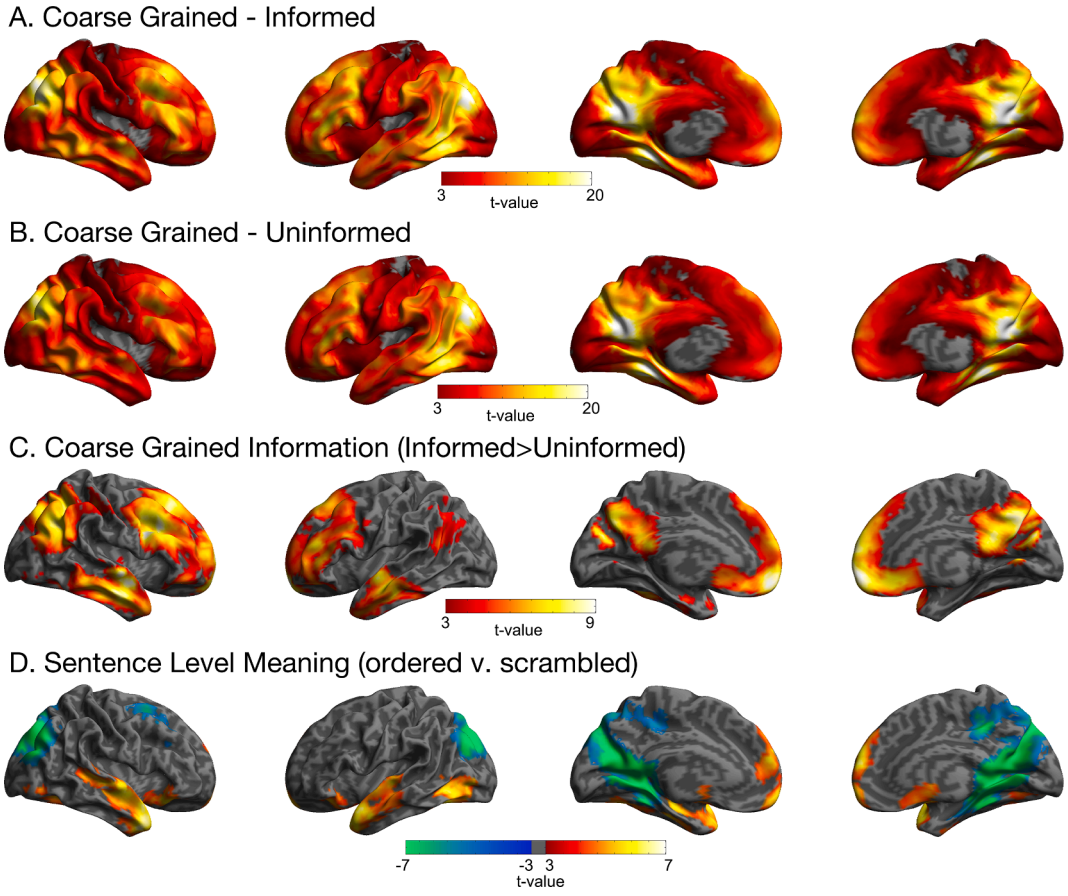


Fig. 3. Results of RSA using the coarse-grained model where RDMs were created by replacing each sentence's individual embedding with the average embedding of the 60 sentences of the thematic category from which that sentence was drawn. **A.** Informed model, created from sentences with canonical word-order. **B.** Binary theoretical model created by coding sentence-pairs as belonging to either the same of different categories **C.** Difference between informed and uniformed brain maps. **D.** Difference in brain maps between informed models created from canonical word-order and word-order scrambled sentences. All brain maps are shown with an initial voxel threshold of $p < .001$, FWE-corrected for multiple comparisons at the cluster level ($p < .05$).

between the brain maps resulting from the informed and uninformed models was determined (Fig. 3C). The primary loci of these course-grained representations are distinct from those seen for the informed and uninformed model, with foci in the anterior superior temporal gyrus (aSTG), vmPFC and the precuneus, posterior parietal lobe, in addition to the lateral PFC (see Table 2). These effects were more pronounced in the right hemisphere, with coarse category level semantic distances more fully capturing continuous inter-category differences, compared to left

Table 2
Peak significance and MNI locations for model comparisons for the coarse-grained (Fig. 3) and fine-grained models (Fig. 4).

	Cluster			Peak T	
Coarse-Grained	**Region**	**P(fwe)**	**Extent**	**t-value**	**xyz**
Informed > Uninformed	raSTS	0.001	57,595	9.93	56 −4 −14
	rIFG_OPE			9.74	42 12 24
	vmPFC			9.66	−10 54 −14
	Prec			8.76	6 −72 30
	posP			8.39	34 −58 32
	lIFG_tri			7.02	−46 38 14
	laSTS	0.001	4698	8.18	−56 −10 −12
	lITG			7.33	−42 −46 −22
Ordered > Scrambled	rTP	0.001	14,812	6.75	44 6 −36
	rOTC			6.52	−44 −60 −12
	lTP			6.32	−50 6 −26
	rOFC			5.89	34 26 −18
	lVTC			5.84	−36 −26 −26
	rAmyg			5.83	20 −4 −8
	mPFC		3423	6.37	4 58 30
	vmPFC			5.76	−8 60 −14
	lOTC		1369	4.96	38 −58 −8
Scrambled > Ordered	lRSC	0.001	23,659	18.14	−16 −62 18
	rRSC			18.06	18 −58 18
	rTOS/AG			11.30	28 −78 36
	rPPG			11.25	26 −40 −12
	lPPG			9.60	−22 −38 −8
	lTOS/AG			9.06	−32 −76 28
	rMFG	0.001	2007	6.74	−28 12 50
Fine-Grained					
Ordered > Scrambled	lIPS	<0.001	2914	5.26	−20 −74 44
	mPFC	<0.001	7693	5.23	−18 28 48
	$lIFG_{OPE}$			5.13	44 16 32
	rMFG			4.86	−40 12 54
	lMTG	<0.001	4030	4.76	−52 −28 −2
	Precuneus	<0.001	1573	4.47	−8 −56 34
	rTOS/AG	<0.001	1280	4.37	34 −76 42
	rpMTG	<0.001	1688	4.36	58 −44 −6
	$lIFG_{ORB}$	0.001	544	4.29	50 28 −14

Abbreviations: IPS: intraparietal sulcus.

hemisphere where category representation appears more absolute in nature.

To isolate the specific contribution of sentence-level meaning, the course-grained model of the ordered sentences was compared to the course-grained model derived from order-scrambled sentences. Fig. 3D shows the clear dissociation of word-order effects across the cortex. Contrary to expectation, we saw that the more nuanced sentence-level information actually impaired the model's ability to capture brain activity in the PPG, TOS/AG, RSC and right MFG (Table 2). This indicates that, at the course level, neural representations in these regions are best captured by the composite words within the sentence, rather than by the full meaning of the sentence. In contrast, sentence-level meaning contributed to coarse-grained category level representations in the anterior temporal poles, ventral temporal cortex, the amygdalae and ventromedial PFC (Table 2).

3.1.2. Fine-grained model

Fine-grained representation of sentence meaning was isolated by repeating the RSA procedure separately within each of the four categories and averaging the results. Brain maps are shown in Fig. 4. Both ordered (Fig. 4A) and scrambled (Fig. 4B) sentence models show maximal information correspondence in regions distinct from both to the full and course-grained models, with maximal information correspondence in left MTG and left IPS.

The added meaning found in sentences was seen to strongly dissociate from that seen at the coarse-grained level, with foci in the inferior and posterior parital lobe, posterior middle temrpoal gryrus extending in the the mid MTG in the left hemisphere, the precuneus and elements of the medial and lateral prefrontal cortex (see Fig. 4C and see Table 2), with only mPFC expressing neuroconceptual convergence with both fine- and coarse-grained models.

3.1.3. Relationship between model grain and information capture

Between category, coarse-grained, differences are expected to dominate the capture of neural representations, as RSA weights distinct differences between representations more than subtle ones. Put another way, informational measures are much more apt at distinguishing between categories, say between apples and oranges, than they are at distinguishing the subtle differences within sets of apples or within sets of oranges. There is simply more information distinguishing sentences drawn from different categories and for this reason, coarse-grained between category differences can be expected to overwhelm fine-grained within-category differences.

This has a counterintuitive influence on our data. The full model, which contains both coarse-grained (apples versus oranges) and fine-grained information (differences between apples and between oranges), actually underperformed compared to coarse-grained models (this is true both for informed and uninformed coarse-grained models – see supplementary Fig. S3). This is counterintuitive as the full model contains more information. However, the result becomes less mysterious when one considers a) how coarse-grained differences dominate the informational space and b) the template model is imperfect. To continue the earlier analogy, if the occasional description of an apple or an orange was wrong for some reason, then one would expect that when distinguishing between apples and oranges replacing the fine descriptions with the generalised coarse description of that set would result in a performance gain. In this way, the extent of within-RDM averaging in this comparison (unlike the main analyses, see methods) is unbalanced and can produce extraneous increases in information capture. This performance gain offsets the lost fine-grained information. This is the most parsimonious, if mundane, explanation for increased information capture for the coarse-grained models compared to the full model.

3.1.4. Reaction time effects

To characterise the influence reaction time (RT) on the main analyses, a control analysis was performed for the full model. The analysis was the same as the main analysis except that, for each subject, an additional RDM was constructed based on RT differences between sentences. This RDM was then included as a regressor of no interest in the semantic searchlight RSA to partial out RT effects. For the a) ordered and b) scrambled models, and c) the difference between ordered and scrambled models, results were highly consistent with the main analysis in terms of distribution, peak location and significance (see supplementary Fig. S2 A-C). A quantitative analysis of the influence of RT differences was accomplished by subtracting the RT-controlled RSA from the RSA of the main analysis for the ordered full model. RT effects were seen to be present in the precentral sulcus, frontal operculum, early visual cortex and the supplementary motor area (see supplementary Fig. S2 D). These regions were distinct from those showing maximal semantic representation. Collectively, these results suggest that RT effects do not contribute meaningfully to the results reported elsewhere in this study.

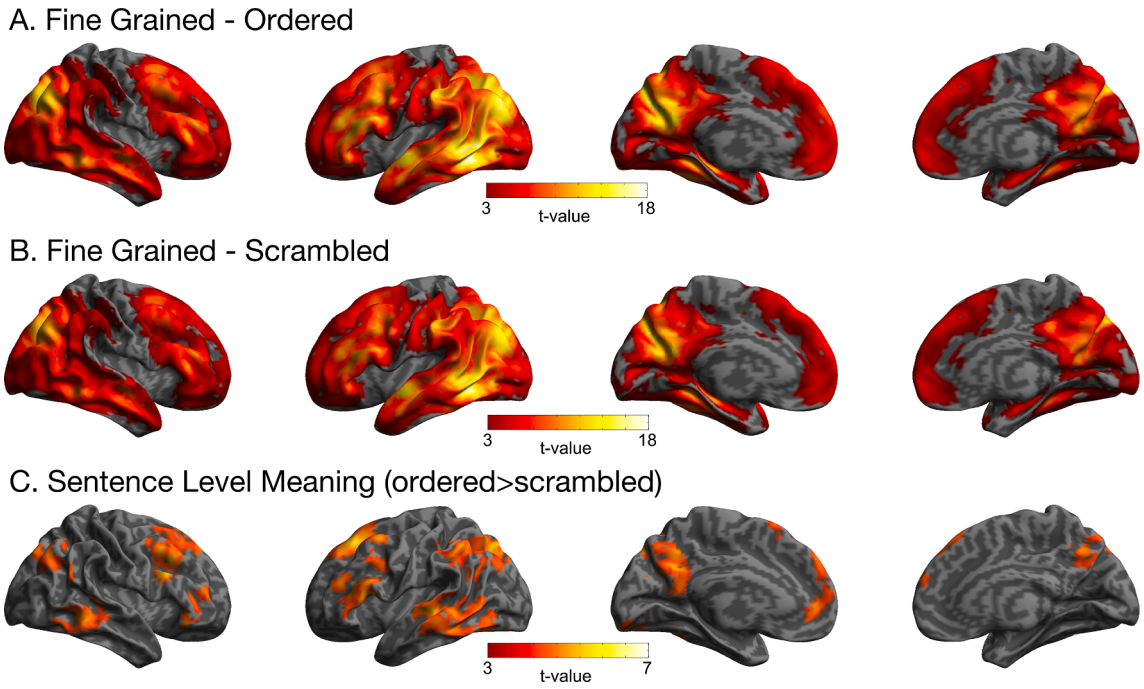


Fig. 4. Results of RSA using the fined-grained model where RSA was conducted separately for each of the four thematic categories and the results averaged **A.** Model created from sentences with words in canonical order. **B.** Model created by word-order scrambled sentences. **C.** Differences between ordered and scrambled-order brain maps. All brain maps are shown with an initial voxel threshold of $p < .001$, FWE-corrected for multiple comparisons at the cluster level ($p < .05$).

4. Discussion

In this work, we used sentence-level embeddings to determine whether the combinatorial meaning contained within sentences relies on neural substrates dissociable from those underlying single-word meaning. When considering the full model, we observed that sentence-level meaning boosted the model's ability to capture information, but did so in a relatively uniform manner across the cortex. In contrast, when broad, course-grained representations and fine-grained representations were considered separately, dissociations became evident in the regions that represented word-level and sentence-level meaning and the regions involved were seen to differ between coarse- and fine-grained representations of meaning.

4.1. Full-model

Examination with the full model revealed the widespread neuro-conceptual convergence for sentence-level embeddings. At the same time, this capacity was largely mirrored by the scrambled model, suggesting that words alone can capture representational spaces across those same brain regions. Indeed, for both ordered and scrambled-order models, representations were seen to peak in pMTG, IFG, anterior STG and left lateral OFC with a left-hemispheric bias was. This pattern is consistent with previous studies of single-word embeddings, both when considered for single word presentation (Liuzzi et al., 2020) and when the average word embeddings computed for words are presented within a sentence (Acunzo et al., 2022). A notable exception to this pattern in the present study was the high correspondence between the full-model and neural representational spaces observed in RSC, TOS/AG and pPG. As noted, these regions are strongly implicated in the representation of place-related words (Fairhall & Caramazza, 2013a, 2013b; Fairhall et al., 2014) and with regard to this specific stimulus set, they are known to exhibit univariate increases in activity for sentences about places (Rabini et al., 2023; Ubaldi et al., 2022). One possibility is that this pattern reflects a response to the presence of place information (see Section 4.2 for further discussion).

The primary question of this study is whether the representation of the additional information associated with combinatorial meaning relies on the same or different neural substrates compared to individual word meaning. There was not robust evidence for such divergence with the full model. When the difference between ordered and scrambled models was considered, the location of the maximal information difference was highly consistent with those exhibiting the maximal information for both ordered and scrambled models. An ROI analysis was performed to further assess the apparent homogeneity of informational capture associated with sentence-level meaning. Here the addition of sentence-level meaning was seen to produce a 20–25 % increase in the capacity of the model to capture neural representation. While this differed between the set of regions previously associated with place selectivity (RSC, PPG, TOS/AG) and those more generally associated with semantic processing (IFG, aSTS, pMTG, SMG, latOFC), it did not differ within this latter subdivision. Thus, at the level of the full model, there do not seem to be specific elements of the semantic system that have a particular importance in combinatorial sentence meaning.

4.2. Coarse-grained model

To examine whether broad differences in meaning, such as that seen between sentences drawn from different thematic categories, are encoded distinctly in the brain, we replaced each sentence's embedding with the average embedding of that category, then compared this model to an uninformed category model that considered the categories to be equidistant. Results demonstrated a distinct change in cortical topography (compare Fig. 2C and 3C). The informed model led to have higher neuroconceptual similarity in the lateral aSTSs, medial PFC and the precuneus, while there was no difference in pMTG. Notably, this network showed a pronounced rightward bias, indicating that course grained differences, as captured by the sentence-embeddings, are mirrored more in the representations of the brain of the right hemisphere than the left. A potential mechanistic explanation for this effect is that the left hemisphere treats categories as discrete entities with the relative difference between categories being of little relevance to computations occurring in these regions. This relative agnosis to coarse-grained differences may result from the specialisation to focus on fine-

grained details associated with the left hemisphere's documented specialisation for semantic meaning (Binder, Desai, Graves, & Conant, 2009). On the other hand, the distinctiveness of categories may be less pronounced in relatively less specialised right hemisphere regions, resulting in a more continuous relationship between categories, with more similar categories being represented in a more similar way. This principle may extend to regions like the pMTG, which show distinct representations of category and the lateral aSTS, which encoded broad relations between them.

The representations discussed in the previous paragraph could reflect the information contained in the constituent words of the sentence or in the sentence's combinatorial meaning. To isolate the added contribution of combinatorial meaning, we again compared ordered to scrambled models. Representations were seen to clearly dissociate across the cortex with coarse-grained sentence-level information being present in the temporal poles, the left lateral occipitotemporal cortex, extending along the ventral temporal cortex anteriorly and in the mPFC. The nature of this effect is complex. It implies that the types of relational information contained within sentences is of particular importance for the representation of broad-scaled differences between these categories of sentences. Dorsal elements of lateral anterior temporal lobe have recently been associated with the language-derived representation of conceptual knowledge (Bi, 2021; Wang et al., 2023; Wang, Men, Gao, Caramazza, & Bi, 2020) and the relational structures within sentences may be more relevant to this form of representation. However, future research will be needed to better understand the exact nature of the relationship of these regions to the coarse-grained meaning afforded by the relational information within sentence structure.

Counter to expectation, representations in place-selective RSC, pPG, TOS/AG and right MFG were captured better by the word-scrambled than the word-ordered model. This indicates that, at the coarse-grained level, the subtle relational meaning conveyed by the sentence structure was irrelevant to (and actually impairs) informational representations in these regions. This provides a further indication that effects in these regions are driven by category selective properties that may not generalise beyond the stimulus set used in this study. It is notable that we see a different pattern regarding the role of these regions when we consider the effects of word order in the full-model (which contain information of both a coarse- and fine-grained nature). In the full-model, ordered models capture more information in classically place-selective regions, RSC, pPG, TOS, while in the coarse model, the reverse is true. This suggests a complex interplay between the contribution of word order to coarse- and fine-grained representations which cannot be fully resolved by the present study.

4.3. Fine-Grained model

Considering overall information captured by both the ordered and scrambled models, the information captured by fine-grained models was more left-lateralised than that captured by either the full or coarse-grained models. Fine-grained representations of meaning peaked in left pMTG, and TOS/AG, a pattern consistent with the specialisation of these left-lateralised regions for semantics (e.g. Binder, 2009) and the consequent ability to make fine grained distinctions between more similar sentences. Additional sentence-level information (ordered > scrambled) again showed a more pronounced left-ward bias and was maximal in IPS, posterior/mid MTG, lateral PFC in both hemispheres along with the precuneus and mPFC. While most of these regions are reliably associated with semantic processing, the contribution of the IPS to sentence-level meaning is uncertain. Meta-analysis has implicated the left IPS in semantic control processes (Noonan, Jefferies, Visser, & Lambon Ralph, 2013), although a more recent *meta*-analysis failed to replicate this result (Jackson, 2021).

A central finding of this study is that, with the exception of the mPFC, all regions showing sensitivity to sentence-level information for fine-grained differences between sentence meaning were distinct from those showing sensitivity at the coarse grain. This indicates that fine-grained specialisation may be associated with a reduced sensitivity to coarse grained information.

4.4. Further considerations

Embeddings are primarily developed to index the meaning of and relatedness between sentences. However, neural networks are black boxes and the underlying mechanisms by which they perform their operations tends to be opaque. For this reason, the nature of the representations captured in this study remains uncertain. For instance, embeddings are sensitive to grammar and syntax and these may potentially play some role in the present result. Future work will benefit from additional models that consider other differences between stimuli that may account for RSA results, such as word frequency or syntax.

At the same time, this study employed naturalistic sentences that conveyed a diversity in meaning. While sentences were matched across knowledge domains on number of words and letters, they were not controlled for all factors (e.g., age of acquisition, familiarity, frequency, orthographic neighbourhood density, ratio of content to functional words etc), which may have potentially affected the results. While one might expect these factors to be balanced within ordered and scrambled models, the influence of these extraneous factors cannot be excluded, especially in so far as they converge with properties with which the embedding model is sensitive. Future studies may benefit from more tightly controlled sentences that are closely matched in linguistic syntactic structure while differing systematically in terms of semantic content. Likewise, future work may benefit from stimulus sets that manipulate the influence of sentence order on meaning ('dog bites man' versus 'man bites dog') to assess both the impact of this factor on sentence-level meaning and the capacity of these large language models to capture such meaning in the brain.

A final consideration is that the four thematic knowledge-domains used in this study may represent an under-sampling the category space, and thus our coarse-grained results reported here may be specific to elements of the four categories used. Future work is needed to ascertain that these findings generalise to coarse-grained representation in general.

While the opaque nature of AI-models can be considered a challenge in studies like the present one. It is also a strength. As AI begins to play a larger role in our day-to-day lives, Explainable/Interpretable AI, and the right to a clear explanation as to why an AI system made the decision it did, is becoming increasingly important to society (White House Office of Science and Technology Policy, 2022). For instance, the sentence-embeddings used in this study are an integral part of chat-GPT. Here the brain can provide a resource to probe the underlying mechanisms of the AI, such as the way in which coarse- and fine-grained sentence meaning that feeds into this large language model maps onto largely distinct human cognitive systems, can contribute to our understanding of how these models perform their tasks.

4.5. Summary

In this work, we used sentence-level word embeddings to gain insight into the representation of combinatorial meaning in the brain. When considered in the context of the undifferentiated full model, that collapsed across coarse- and fine-grains, while the capacity of sentence embeddings to capture neural representational spaces was largely contingent on the specific words irrespective of sentences structure it was seen to increase by ~ 20–25 % when information about sentence structure was conserved, in a manner that was relatively uniformly across brain regions. However, when divided into coarse- and fine-grained distinctions in sentence meaning, differences were observed in dissociable brain regions. Collectively, these results indicate that differing neural systems are biased towards single-word and combinatorial meaning and additionally that the brain is organised into cortical

systems more specialised for fine-grained or coarse-grained meaning, where category boundaries are more discrete and absolute in regions specialised for fine-grained meaning and blurred in less specialised regions allowing the representation of broad scale relationships between dissimilar sentences.

CRediT authorship contribution statement

Scott L. Fairhall: Conceptualization, Formal analysis, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Silvia Ubaldi and Giuseppe Rabini for their help with stimulus development and data collection for the datasets used in this study. The project was funded by the European Research Council (ERC) grant CRASK - Cortical Representation of Abstract Semantic Knowledge, under the European Union's Horizon 2020 research and innovation program (grant agreement no. 640594). MRI scanning was supported by funding from the Caritro Foundation, Italy.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2024.105389.

References

Acunzo, D. J., Low, D. M., & Fairhall, S. L. (2022). Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. *NeuroImage, 251*(February), Article 119005. https://doi.org/10.1016/j.neuroimage.2022.119005

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., … Lalor, E. C. (2021). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience, 41* (18), 4100–4119.

Anderson, A. J., Lalor, E. C., Lin, F., Binder, J. R., Fernandino, L., Humphries, C. J., … Wang, X. (2019). Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. *Cerebral Cortex, 29*(6), 2396–2411. https://doi.org/10.1093/cercor/bhy110

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences, 25*(10), 883–895.

Bi, Y., Wang, X., & Caramazza, A. (2016). Object Domain and Modality in the Ventral Visual Pathway. *Trends in Cognitive Sciences, 20*(4), 282–290. https://doi.org/10.1016/j.tics.2016.02.002

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex, 19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Bruffaerts, R., Dupont, P., Peeters, R., De Deyne, S., Storms, G., & Vandenberghe, R. (2013). Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33*(47), 18597–18607. https://doi.org/10.1523/JNEUROSCI.1548-13.2013

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5*(1). https://doi.org/10.1038/s42003-022-03036-1

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*. https://doi.org/10.1038/13217

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33*(48), 18906–18916. https://doi.org/10.1523/JNEUROSCI.3809-13.2013

Epstein, R. A., Parker, W. E., & Feiler, A. M. (2007). Where Am I Now? Distinct Roles for Parahippocampal and Retrosplenial Cortices in Place Recognition. *Journal of Neuroscience, 27*(23), 6141–6149. https://doi.org/10.1523/JNEUROSCI.0799-07.2007

Epstein, R., & Kanwisher, N. (1998). A cortical representation the local visual environment. *Nature*. https://doi.org/10.1038/33402

Fairhall, S. L., & Caramazza, A. (2013a). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33*(25). https://doi.org/10.1523/JNEUROSCI.0051-13.2013

Fairhall, S. L., & Caramazza, A. (2013b). Category-selective neural substrates for person- and place-related concepts. *Cortex, 49*(10), 2748–2757. https://doi.org/10.1016/j.cortex.2013.05.010

Fairhall, S. L. (2020). Cross recruitment of domain-selective cortical representations enables flexible semantic knowledge. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2224-19.2020

Fairhall, S. L., Anzellotti, S., Ubaldi, S., & Caramazza, A. (2014). Person- and place-selective neural substrates for entity-specific semantic access. *Cerebral Cortex, 24*(7), 1687–1696. https://doi.org/10.1093/cercor/bht039

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., … Seidenberg, M. S. (2015). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex, 1–17*. https://doi.org/10.1093/cercor/bhv020

Friston, K. J., Rotshtein, P., Geng, J. J., Sterzer, P., & Henson, R. N. (2006). A critique of functional localisers. *NeuroImage, 30*(4), 1077–1087. https://doi.org/10.1016/j.neuroimage.2005.08.012

Fu, Z., Wang, X., Wang, X., Yang, H., Wang, J., Wei, T., … Bi, Y. (2023). Different computational relations in language are captured by distinct brain systems. *Cerebral Cortex (New York, N.Y.: 1991)*, *33*(4), 997–1013. https://doi.org/10.1093/cercor/bhac117.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences, 10*(2), 64–69. https://doi.org/10.1016/j.tics.2005.12.005

Jackson, R. L. (2021). The neural correlates of semantic control revisited. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2020.117444

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America, 103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103.

Leonardelli, E., & Fairhall, S. L. (2022). Similarity-based fMRI-MEG fusion reveals hierarchical organisation within the brain's semantic system. *NeuroImage, 259* (June), Article 119405. https://doi.org/10.1016/j.neuroimage.2022.119405

Liuzzi, A. G., Aglinskas, A., & Fairhall, S. L. (2020). General and feature-based semantic representations in the semantic network. *Scientific Reports*. https://doi.org/10.1038/s41598-020-65906-0

Liuzzi, A. G., Bruffaerts, R., Dupont, P., Adamczuk, K., Peeters, R., De Deyne, S., … Vandenberghe, R. (2015). Left perirhinal cortex codes for similarity in meaning between written words: Comparison with auditory word input. *Neuropsychologia, 76*, 4–16. https://doi.org/10.1016/j.neuropsychologia.2015.03.016

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., … Weng, L. (2022). Text and Code Embeddings by Contrastive Pre-Training. Retrieved from http://arxiv.org/abs/2201.10005.

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond inferior prefrontal involvement in semantic control: Evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *Journal of Cognitive Neuroscience, 25*(11), 1824–1850. https://doi.org/10.1162/jocn_a_00442

Noppeney, U., Price, C. J., Penny, W. D., & Friston, K. J. (2006). Two distinct neural mechanisms for category-selective responses. *Cerebral Cortex, 16*(3), 437–445. https://doi.org/10.1093/cercor/bhi123

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in matlab/GNU octave. *Frontiers in Neuroinformatics*. https://doi.org/10.3389/fninf.2016.00027

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., … Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications, 9*(1). https://doi.org/10.1038/s41467-018-03068-4

Rabini, G., Ubaldi, S., & Fairhall, S. L. (2021). Combining Concepts Across Categorical Domains: A Linking Role of the Precuneus. *Neurobiology of Language, 2*(3), 354–371. https://doi.org/10.1162/nol_a_00039

Rabini, G., Ubaldi, S., & Fairhall, S. L. (2023). Task-based activation and resting-state connectivity predict individual differences in semantic capacity for complex semantic knowledge. *Communications Biology, 6*(1020), 1–13. https://doi.org/10.1038/s42003-023-05400-1

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America, 118*(45). https://doi.org/10.1073/pnas.2105646118.

Silson, E. H., Steel, A. D., & Baker, C. I. (2016). Scene-selectivity and retinotopy in medial parietal cortex. *Frontiers in Human Neuroscience, 10*(August), 17. https://doi.org/10.3389/fnhum.2016.00412

Simanova, I., Hagoort, P., Oostenveld, R., & Van Gerven, M. A. J. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhs324

Sun, J., Wang, S., Zhang, J., & Zong, C. (2021). Neural Encoding and Decoding with Distributed Sentence Representations. *IEEE Transactions on Neural Networks and Learning Systems, 32*(2), 589–603. https://doi.org/10.1109/TNNLS.2020.3027595

Ubaldi, S., Rabin, G., & Fairhall, S. L. (2022). Recruitment of control and representational components of the semantic system during successful and unsuccessful access to complex factual knowledge. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2485-21.2022

Wang, X., Wang, B., & Bi, Y. (2023). Early language exposure affects neural mechanisms of semantic representations. *ELife, 12*, e81681.

Wang, X., Men, W., Gao, J., Caramazza, A., & Bi, Y. (2020). Two forms of knowledge representations in the human brain. *Neuron, 107*(2), 383–393.

White House Office of Science and Technology Policy, X. (2022). *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.* Retrieved from https://www.whitehouse.gov/ostp/ai-bill-of-rights/.