PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Modelling brain representations of abstract concepts

Daniel Kaiser[1,2,3]*, **Arthur M. Jacobs**[4,5], **Radoslaw M. Cichy**[4,6,7]

1 Mathematical Institute, Department of Mathematics and Computer Science, Physics, Geography, Justus-Liebig-Universität Gießen, Gießen, Germany, **2** Center for Mind, Brain and Behavior (CMBB), Philipps-Universität Marburg and Justus-Liebig-Universität Gießen, Marburg, Germany, **3** Department of Psychology, University of York, York, United Kingdom, **4** Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany, **5** Center for Cognitive Neuroscience Berlin, Freie Universität Berlin, Berlin, Germany, **6** Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany, **7** Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

* danielkaiser.net@gmail.com

Citation: Kaiser D, Jacobs AM, Cichy RM (2022) Modelling brain representations of abstract concepts. PLoS Comput Biol 18(2): e1009837. https://doi.org/10.1371/journal.pcbi.1009837

Editor: Leyla Isik, MIT, UNITED STATES

Received: June 2, 2021

Accepted: January 14, 2022

Published: February 4, 2022

Data Availability Statement: Data and code are publicly available on OSF (https://doi.org/10.17605/OSF.IO/FTBJQ).

Funding: D.K. and R.M.C. are supported by DFG grants (KA4683/2-1, CI241/1-1, CI241/3-1, CI241/7-1). R.M.C. is supported by an ERC Starting Grant (ERC-2018-StG 803370). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Abstract conceptual representations are critical for human cognition. Despite their importance, key properties of these representations remain poorly understood. Here, we used computational models of distributional semantics to predict multivariate fMRI activity patterns during the activation and contextualization of abstract concepts. We devised a task in which participants had to embed abstract nouns into a story that they developed around a given background context. We found that representations in inferior parietal cortex were predicted by concept similarities emerging in models of distributional semantics. By constructing different model families, we reveal the models' learning trajectories and delineate how abstract and concrete training materials contribute to the formation of brain-like representations. These results inform theories about the format and emergence of abstract conceptual representations in the human brain.

Author summary

How do we conceive abstract concepts, like love, peace, or truth? In this study, we investigate how our brains support the activation and contextualization of such abstract concepts. We asked participants to embed abstract nouns into a coherent story while we recorded functional MRI. Using multivariate analysis techniques, we computed how similar different abstract concepts were represented during this task. We then modelled these neural similarities among concepts with computational models of distributional semantics which capture the words' co-occurance statistics in large natural language corpora. Our results reveal a correspondence between the computational models and brain representations in the inferior parietal cortex. This correspondence held when the computational models were only trained on subsets of the corpora that contained as few as 100,000 sentences and only abstract or concrete words. Our findings establish a neural correlate of abstract concept representation in the inferior parietal cortex, and they provide a first characterization of the format of these representations.

Introduction

The use of conceptual knowledge is one of the foundations of human intelligence. On the neural level, concepts are represented in a complex network of brain regions [1]. Fueled by novel computational models of distributional semantics, researchers have recently started to unravel the format of concept representations in this neural network. By harnessing linguistic co-occurrence statistics, these models not only capture representations of concepts from written and spoken language [2–5], but also predict representations of novel concepts [6].

However, these recent advances in understanding the representations of conceptual knowledge largely hinge on the study of concrete concepts. And although some of the previous studies (e.g., [3,6]) have included both concrete and abstract words, they probed representations across the two, and therefore, could not reveal how specifically abstract concepts are represented. Only few studies have explicitly investigated how abstract concept representations are organized [7–10]. To date, key questions about the emergence and the format of these representations remain heavily debated [11].

Here, we model brain representations that support the activation and contextualization of abstract concepts. We recorded fMRI while participants were tasked with embedding abstract nouns into a background context. By relating brain activations during this task to targeted models of distributional semantics, we shine a new light on the format of abstract concept representations in the human brain.

Results and discussion

In an fMRI experiment, we visually presented 61 abstract German nouns (see Materials and Methods). Participants (n = 19) read these words and silently embedded them into a coherent story that they were developing around a prespecified contextual background (Fig 1A). This task was chosen to be engaging while ensuring a sufficiently deep level of processing. Here, participants were required to retrieve the meanings of the words and use those meanings, integrating them with a complex ongoing stream of thought.

Cortical responses were modelled in a representational similarity analysis (RSA) framework. In RSA, the representational organization found in the brain is directly compared to the representational organization found in candidate representational models: this is done in a computationally straightforward way, whereby pairwise similarity relations are correlated across a larger set of stimuli [12]. Here, we used an RSA searchlight approach, in which we extracted similarity relations among the words across the whole cortex (Fig 1B; see Materials and Methods). We modelled these similarities using a word2vec model of distributional semantics [13], trained on a 45-million sentence corpus (SdeWaC [14]).

The model predicted brain activations in the left inferior parietal cortex (IPC; Fig 1C), most prominently covering the angular gyrus, but extending into the superior parietal and middle occipital cortices (158 voxels, peak: -36/-58/46, t[18] = 6.16), and in right IPC, with an anterior cluster primarily in the supramarginal gyrus (91 voxels, peak: 45/-46/58, t[18] = 4.91), and a posterior cluster in the angular gyrus (35 voxels, peak: 36/-73/40, t[18] = 4.80). Detailed analyses, reported in the Supplementary Information (Fig B in S1 Text), show that this correspondence was not driven by individual words present in the stimulus set. These results show that bilateral IPC represents semantic similarities of abstract concepts. They further suggest IPC as a cortical hub for the activation and contextualization of abstract concepts.

This interpretation, however, warrants a note of caution: Although it is tempting to strongly interpret the IPC representations as a reflection of concept coding, they may also reflect linguistic coding. While the distributional information extracted from word2vec inherently

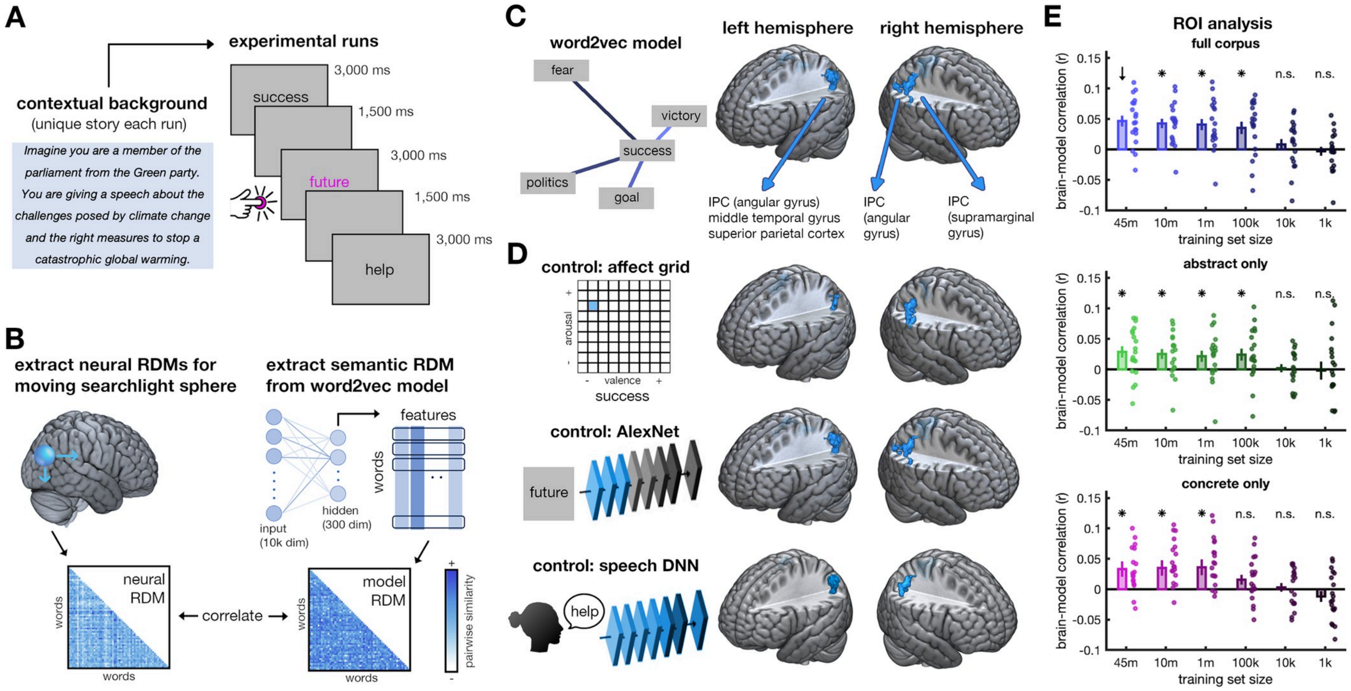

Fig 1. Representation of abstract concepts in parietal cortex. A) Participants completed 10 runs of fMRI recordings. Before each run, they established a unique background context in their mind and were then asked to silently narrate a story of their own making that included the subsequent 61 abstract words, which were presented in a different random order in every run (all German word stimuli in Table A in S1 Text). **B)** In a searchlight analysis, representational dissimilarity matrices (RDMs) were extracted (i) from the brain data, by pairwise cross-validated correlations among localized activity patterns, and (ii) from a word2vec model of distributional semantics, by pairwise correlations among hidden-layer activations. **C)** Correlating the neural and model RDMs revealed clusters in bilateral inferior parietal cortex (IPC), primarily covering the angular gyrus. Brain maps are thresholded at $p_{voxel}<0.001$ (uncorrected) and $p_{cluster}<0.05$ (FWE-corrected). Cross-sectional images of the significant clusters as well as unthresholded statistical maps can be found in the Supplementary Information (Figs G and H in S1 Text). **D)** These clusters persisted when repeating the analyses while partialing out the effects of emotional word content (using affect grids), visual wordform (using a visual-categorization DNN), and auditory properties of the spoken words (using a speech-recognition DNN). **E)** Within the IPC cluster defined on the full 45-million sentence model (marked by an arrow), we compared model families trained on different corpus sizes and on only abstract or concrete words, respectively. Brain-like representations emerged in models that were trained on as little as 100,000 sentences and on either abstract or concrete embeddings. Dots show individual-participant data, error bars denote SEM, asterisks represent $p<0.05$ (FDR-corrected).

https://doi.org/10.1371/journal.pcbi.1009837.g001

contains conceptual properties (e.g., [15–17]), it is primarily linguistic. As also our experimental design used words to activate representations in the brain, the similarities between the model and the brain we see may be driven by the linguistic, rather than the conceptual information these words carry.

Controlling for emotional and sensory word properties

Some researchers have argued that abstract concepts are represented through their grounding in the emotional domain [18,19]. To test whether IPC representations are indeed driven by the words' emotional content, we re-performed the analysis while partialing out valence and arousal ratings (see Materials and Methods). We still found clusters in left (36 voxels, peak: -33/-64/31, t[18] = 4.73), and right (108 voxels, peak: 63/-40/31, t[18] = 7.30) IPC, suggesting that the emotional content is insufficient to explain abstract concept representation in parietal cortex. Notably, the left IPC cluster somewhat shrunk after controlling for emotional word properties. This may be because word2vec models can pick up on emotional features implicitly contained in word embeddings [17].

IPC is also sensitive to sensory properties, such as visual form [20] and phonological speech attributes [21]. However, when repeating the analysis while partialing out early activations in a visual-categorization deep neural network (DNN; see Materials and Methods), we still found clusters in the left (157 voxels, peak: -36/-58/46, t[18] = 6.07) and right (anterior: 76 voxels, peak: 51/-49/52, t[18] = 4.70; posterior: 32 voxels, peak: 39/-76/40, t[18] = 4.82) IPC. Similarly, when controlling for activations obtained from a speech-recognition DNN (see Materials and Methods), we still found clusters in left (113 voxels, peak: -33/-64/31, t[18] = 5.43, $p_{cluster}<0.001$), and right (53 voxels, peak: 48/-46/55, t[18] = 4.68) IPC. These results show that sensory properties are unrelated to IPC representations of abstract concepts.

In the Supplementary Information (Fig D in S1 Text), we additionally show that word frequency cannot account for the correspondence between the word2vec model and brain representations in IPC.

Trajectories towards brain-like representations

Our observation that the word2vec model and IPC share abstract concept representations led us to ask how the model acquires this property. To test whether co-occurrence statistics are acquired incrementally over increasing experience with human language, we devised a word2vec model family whose members were trained on staggered amounts of data, from the full 45-million sentence corpus down to fragments as small as 1,000 sentences. We then evaluated how well models trained on less data could still predict representations in the IPC cluster that yielded the best correspondence with the full 45-million sentence model (see Materials and Methods).

This analysis revealed decreasing correspondence with decreasing training data (mean r = 0.74, t[18] = 5.89, p<0.001). Nonetheless, a model trained on only 100,000 sentences (~0.2% of the corpus) still predicted IPC representations well (t[18] = 3.46, $p_{FDR} = 0.002$; comparison to full model: t[18] = 2.15, p = 0.045), whereas models trained on smaller training sets did not (Fig 1D). Direct comparisons of all models to the full model trained on 45m sentences can be found in the Supplementary Information (Fig E in S1 Text). These results show that brain-like representations are learned through linguistic co-occurrence statistics, which can emerge already from a (relatively) modest degree of training experience.

Modelling brain representations from abstract and concrete embeddings

Some theorists argue that the meaning of abstract concepts needs to be derived through the activation of related concrete concepts, which are in turn grounded in sensory experiences [22,23]. This view prompts the hypothesis that representations of abstract concepts originate primarily from co-occurrence statistics between abstract and concrete words, rather than among abstract words alone. To test this hypothesis, we trained word2vec models on subsets of the 45-million sentence corpus that we devised to consist of abstract or concrete words only (see Materials and Methods).

Models trained on abstract-only and concrete-only corpora both predicted representations in IPC (Fig 1E). Reproducing the previous pattern of results, we found that models trained on larger fractions of the corpus predicted representations better (abstract only: mean r = 0.36, t[18] = 2.55, p = 0.010; concrete-only: mean r = 0.73, t[18] = 4.32, p<0.001). Interestingly, representations were modelled equally well by the most extensively trained abstract-only (t[18] = 3.09, $p_{FDR} = 0.010$) and concrete-only models (t[18] = 2.60, $p_{FDR} = 0.018$; comparison: t[18] = 0.50, p = 0.62), suggesting that brain-like representations of abstract concepts can emerge from either abstract or concrete semantic embeddings.

Conclusions

Our findings yield multiple key insights into abstract concept representation:

First, our findings provide novel evidence that the IPC is a core area for concept coding [1]. Returning to the question of whether our results reflect genuine conceptual representation or language-specific codes, the localization of effects to the IPC indeed provides tentative evidence for the former: IPC activations are not routinely observed in language tasks (see [24])–by contrast, particularly the angular gyrus is often implicated in brain networks for concept representation [25]. Others, however, have recently contested this role of the IPC, as the region is not consistently activated during semantic cognition [26,27]. Critically, the current study used a task that required participants to activate and contextualize abstract concepts. In this task, we identify the angular gyrus as a critical hub for the dynamic use of abstract knowledge, consistent with the view that this region plays a key role in combinatory linguistic processing [28–32]. Such combinatory processing may be particularly critical for abstract concepts, which more strongly need to be contextualized in a situational way during everyday use. It is worth noting, however, that our study does not establish that the angular gyrus is specifically important for representing abstract but not concrete words. In fact, previous results suggest that concrete words activate the angular gyrus just as strongly as abstract words [33], so that future studies need to carefully compare the representation of abstract and concrete concepts in this region.

Second, our study shows that brain representations of abstract concepts can be predicted from distributional word embeddings in natural language [10]. Interestingly, the organization of abstract concepts, as found in our brains, can be modelled from linguistic embeddings in both abstract and concrete realms of knowledge. This result shows that despite their representational dissimilarities [34,35], abstract and concrete concepts may be organized through shared principles. It is worth noting that the models constructed from abstract-only and concrete-only corpora in our study still produced a moderately high inter-correlation (r = 0.78). Although this suggests a similarity in abstract words' embeddings within other abstract and concrete words, this high correlation also limits the potential of our analysis to reveal substantial differences in how well these models predict neural representations. Future studies could specifically assemble stimulus sets that target concepts for which embeddings in the abstract and concrete realms are more different.

Third, our data informs theories of abstract knowledge representation [11]. Our results do not provide evidence for theories suggesting that abstract concepts are coded solely through emotional associations [19] or the activation of related concrete concepts [23,36]. Further, in our study, we did not find evidence for an additional visual representation of abstract concepts [7,37] or for a grounding of abstract conceptual knowledge in cognitive/motor systems [38]. Our findings rather suggest that abstract knowledge is reflected in distributional relationships in neural representations of the concept or language processing systems. However, the question how distributional codes (such as the ones capitalized on by language models like word2vec) relate to word meaning is controversial: positions range from claims that word meaning is determined by distributed relations and respective neural codes akin to those in word2vec [39–41] to the argument that distributional codes are insufficient to provide insights into meaning [42]–under this view, the observed similarities might be an emerging phenomenon rather than the underlying coding scheme. The current investigation cannot arbitrate between such theoretical positions.

Fourth, our results suggest that by harnessing co-occurrence statistics from linguistic experience, computational models of distributional semantics can acquire abstract concept representations that are organized in similar ways as biological representations. Although massive corpora are immensely popular for modelling language organization, our analyses of model learning trajectories show that brain-like representations can emerge from much smaller training sets of only 100,000 sentences. Cleary, our study provides only a first, coarse approximation of the tentative learning trajectory towards brain-like representations. Future work needs to map out the emergence of more fine-grained information along these learning trajectories to investigate how closely the acquisition of the models' representations across training can predict human concept learning and development [43].

Finally, our study highlights that computational models–through systematic manipulation of model training regimes–can yield targeted insights into the emergence and format of concept representations. Moving ahead, future studies could not only refine training regimes but also comprehensively manipulate a set of fundamental model parameters [44]. First advances have recently been made by comparing language models with different architectures to brain data [45], by enriching models with predictive and contextual information [46–49] and by testing the applicability of linguistic models to experiences in domains like vision [50–52]. In the future, employing such targeted model-based analyses may yield further fine-grained insights into how our brain represents abstract knowledge.

Materials and methods

Ethics statement

All procedures were approved by the ethical committee of the Department of Education and Psychology, Freie Universität Berlin, and were in accordance with the Declaration of Helsinki. Formal written consent was obtained from all participants.

Participants

Nineteen healthy adults (mean age 28.8 years, SD = 6.1; 10 female) with normal or corrected-to-normal vision completed the experiment. All of them were right-handed and native German speakers. Participants were recruited from the online participant database of the Berlin School of Mind and Brain [53] and received monetary reimbursement or course credits.

Stimuli and paradigm

The stimulus set consisted of 61 abstract German nouns. These nouns were chosen from a list of the most frequent German words (from: wortschatz.uni-leipzig.de), from which they were arbitrary selected to cover a range of themes. All words and their English translations can be found in the Supplementary Information (Table A in S1 Text).

During the fMRI experiment, participants completed 10 runs. Before each run, participants read through one of 10 contextual background stories. All texts and their English translations can be found in the Supplementary Information (Table B in S1 Text). Participants were asked to mentally image themselves being in the scenario outlined in the text. After reading through the text, participants were instructed to use the subsequently presented words in the upcoming run to mentally narrate a story that incorporates the words as they are shown on the screen. They were instructed that it is completely up to them how the story unfolds as long as they use all the words in their story. Stories were chosen to be emotionally engaging to increase participants' immersion into the task. The order of the 10 stories was randomized for every participant.

Each run contained 61 experimental trials. On each trial, one of the abstract words was shown for 3 seconds, in black Arial font on a gray background. Trials were separated by an inter-trial interval of 1.5 seconds, during which a fixation cross was shown. In addition to the experimental trials, each run included 14 fixation trials, where only the fixation cross was shown throughout the trial. Trial order was randomized within each run.

To ensure that participants paid attention to the words, we introduced a simple manual task: In each run, 7 of the word were shown in pink color and participants had to press a button whenever they saw one them.

Runs started and ended with brief fixation periods; each run lasted 5:48 minutes. The stimulation was back-projected onto a translucent screen at the end of the MRI scanner bore and controlled using the Psychtoolbox [54].

Additionally, prior to the experiment, each participant completed a practice run (using a background text different from the ones used in the experiment).

Two participants completed a version of the experiment that differed in two aspects: the inter-trial interval was 1s instead of 1.5s and no behavioral task was included.

MRI acquisition and preprocessing

MRI data was acquired using a 3T Siemens Tim Trio Scanner equipped with a 12-channel head coil. T2*-weighted gradient-echo echo-planar images were collected as functional volumes (TR = 2s, TE = 30ms, 70° flip angle, 3mm^3 voxel size, 37 slices, 20% gap, 192mm FOV, 64×64 matrix size, interleaved acquisition). Additionally, a T1-weighted image (MPRAGE; 1mm^3 voxel size) was obtained as a high-resolution anatomical reference. Preprocessing was done in MATLAB using SPM12 (www.fil.ion.ucl.ac.uk/spm/). The functional volumes were realigned and coregistered to the T1 image. The T1 image was normalized to MNI-305 standard space to obtain transformation parameters used to normalize participant-specific results maps (see below).

Representational similarity analysis

To quantify neural representations, we used multivariate representational similarity analysis (RSA) [12]. In RSA, neural representations are first characterized by means of their pairwise similarity structure (i.e., how similarly each stimulus is represented with each other stimulus). The pairwise dissimilarities between neural representations are organized in neural representational dissimilarity matrices (RDMs) indexed in rows and columns by the experimental conditions compared. Then, the neural similarity structure (i.e., the neural RDMs) are correlated to model RDMs, which capture different aspects of the conditions' similarity. Significant correlations between the neural RDMs and these model RDMs indicate that the aspect of similarity conveyed by the model is represented in the brain.

In recent studies, RSA has successfully been used to relate representations in computational models of language processing and human cortex [10,49,55,56]. Other studies have probed correspondences between language processing models and the brain through the use of encoding models [2,3,6,48,50]. Encoding models seek to directly establish a mapping between the feature dimensions extracted by the computational model and fMRI responses in individual voxels. For this task, they require diverse and large sets of data to train the model weights [57], which the current design was not optimized for. RSA and encoding models offer largely complimentary quantifications of neural representations, with comparable sensitivity [58] and comparable constraints regarding interpretability [59]. However, although RSA offers a straightforward and computationally efficient way to relate computational models and population codes in the brain, one limitation needs to be taken into account when interpreting the results: representations in some parts of the brain may rely on intricate weightings of few feature dimensions and may therefore be harder to identify with RSA than with encoding models.

Extracting neural dissimilarity. Separately for each participant and each run, we first modeled the functional MRI data in a general linear model (GLM) with 67 predictors (61 predictors for the 61 words, and 6 predictors for the 6 movement regressors obtained during realignment). From these GLMs, we obtained 610 beta weights of interest for every voxel, which quantified the voxel's activation to each of the 61 words in each of the 10 runs. All further analyses were carried out using a searchlight approach [60], that is, analyses were done repeatedly for a spherical neighborhood (3-voxel radius) centered on each voxel across the brain. This approach allowed us to quantify and model neural representations in a continuous and unconstrained way across brain space.

For each searchlight neighborhood, neural RDMs were created based on the similarity of multi-voxel response patterns, using the CoSMoMVPA toolbox [61]. Within each neighborhood, we extracted the response pattern across voxels evoked by each word in each run. We then performed a cross-validated correlation analysis [62]. Unbiased, cross-validated distance metrics like cross-validated correlations are generally considered more reliable than non-cross-validated metrics (such as plain correlations) for estimating pattern similarities in brain data [63]. For this analysis, the data were repeatedly split into two halves (all possible 50/50 splits; results were later averaged across these splits) and the response patterns for each word were averaged within each half. For each pair of words, we then computed two correlations: (i) within-condition correlations were computed by correlating the response patterns evoked by each of the two words in one half of the data with the response patterns evoked by the same word in the other half of the data, and (ii) between-condition correlations were computed by correlating the response patterns evoked by each of the two words in one half of the data with the response patterns evoked by the other word in the other half of the data. By subtracting the between-correlations from the within-correlations for each pair of words, we obtained an index of how dissimilar two words are based on the response patterns they evoked in the current searchlight neighborhood. Repeating this analysis for each pair of words yielded a 61×61 neural RDM for each searchlight.

Modelling neural dissimilarity. To model the semantic representation of the abstract words, we used a word2vec computational model of distributional semantics [13]. The model was trained on the SdeWaC corpus, which contains 45 million German sentences [14], using the gensim library (https://github.com/RaRe-Technologies/gensim). The model hyperparameters were the following: dimensions = 300, model type = skipgram, windowsize = 5, minimum count = 1, iterations = 50. For each word in the corpus, this model yields a vector representation that indicates its position in a 300-dimensional vector space. Distances in this vector space reflect similarities in word embeddings. We then created a 61×61 RDM based on the pairwise correlations of the 300 vector-space features for each of the words used in the experiment.

To establish correspondences between the model and the brain data, the model RDMs were correlated with the neural RDMs for each searchlight, using the lower-off diagnonal entries of each RDM. These correlations were then Fisher-transformed and mapped back to the searchlight center. We thereby obtained brain maps of correspondence between each model and the neural data. For each participant, these maps were warped into standard space by using the normalization parameters obtained during preprocessing.

Controlling for emotional, visual, and auditory word properties. As an emotional content model, we used participants' responses in an affect grid task, where 20 participants (partly including the participants in the current experiment) concurrently rated each word's valence and arousal by selecting one compartment of a 9×9 grid [64]. From these data, we created two RDMs: (i) a valence RDM, whose entries reflected pairwise absolute difference in the words' valence ratings and (ii) an arousal RDM, whose entries reflected pairwise absolute difference in the words' arousal ratings. The valence and arousal RDMs were mildly correlated with each other ($r = 0.19$) and with the different word2vec model RDMs (all $r<0.24$). The words' similarity in valence an arousal did not significantly predict brain activations in a searchlight analysis.

As a visual word form model, we used activations in the three earliest convolutional layers an AlexNet DNN pre-trained on object recognition [65,66], which have been shown to capture representations of simple visual attributes in visual cortex [67]. We printed the 61 words as they appeared in the experiment on a 225×225 pixel gray image background and fed these images to the DNN. The resulting network activations were used to construct model RDMs. For each of the first three convolutional layers of the network, the RDM was constructed by computing pairwise distances (1-correlation) between layer-specific activation vectors. The visual DNN RDMs were only very weakly correlated with the word2vec model RDMs (all r<0.1). Searchlight analyses revealed that the first three layers of the visual DNN predicted activations in bilateral posterior visual cortex, including fusiform cortex (see Fig A in S1 Text).

As a model of auditory, phonetic word similarity, we used activations in a DNN model of auditory speech recognition [68]. We obtained spoken versions of the 61 words from the ttsmp3 webpage (https://ttsmp3.com/text-to-speech/German/). The sound files were resized to a length of 2 seconds by right-padding them with zeros, transformed into a cochleagram representation, and then passed through the speech recognition branch of the DNN. The resulting network activations were used to construct model RDMs. For each of the seven layers of the network, the RDM was constructed by computing pairwise distances (1-correlation) between layer-specific activation vectors. The auditory DNN RDMs were only weakly correlated with the word2vec model RDMs (all r<0.16). Searchlight analyses revealed that the early layers of the auditory DNN, because of the correlation between word length and speech duration, also predicted activations in bilateral posterior visual cortex. By contrast, the last layer of the network specifically predicted activations in left middle temporal gyrus (Fig A in S1 Text).

To control for emotional and sensory properties, we performed searchlight analyses relating the neural RDMs and the word2vec model RDMs as before, while we partialed out the two emotion predictor RDMs, the three visual DNN predictor RDMs, or the seven auditory DNN predictor RDMs, respectively. Specifically, for each searchlight neighborhood, we computed a partial correlation between the neural RDM and the predictor RDM which was controlled for the control RDMs. This procedure ensured that if the control RDMs predicted the same portion of variance in the neural RDM as the predictor RDM, the correlation would disappear (for similar approaches, see [69–71]). All other aspects of the analysis remained identical to the previous searchlight analysis.

Region of interest analyses. For further dissecting the representations in left parietal cortex, we specifically focused on this area in a region-of-interest (ROI) analysis. The IPC clusters that showed significant correspondence with the word2vec model in the main analysis were chosen as the ROI. Neural RDMs were generated from pairwise correlations of activity patterns across all voxels in the ROI; the procedure was otherwise identical to the procedure applied in the searchlight analysis (see above).

As ROI definition was done on the basis of the model that was trained on the full 45-million sentences SDeWaC corpus, we never evaluated this model statistically in our ROI analysis. We instead probed the correspondence between neural RDMs in the ROI and RDMs built from a set of different word2vec model families whose training regimes differed in important aspects.

To probe the behavior of our word2vec model with changes in training set, we created a model family whose members were trained on different amounts of data. Models were trained on different fragments of the corpus (containing 45m, 10m, 1m, 100k, 10k, or 1k sentences). Each of these fragments corresponded to the first *n* sentences in the corpus (e.g., the 1k model comprised the first 1,000 sentences). We thereby ensured that the smaller fragments were always completely included in the larger ones.

Additionally, we constructed a model family whose members were trained on abstract words and a model family whose members were trained on concrete words. Members in each family differed by the amount of data they were trained on (as outlined above). Abstract and concrete words were defined on the basis of the abstractness-concreteness scale of the IMS norms (https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/affective-norms) [72]. For the abstract-only models, we chose words that had a z-value of <0 on the on this scale and removed all other words from the corpus; this left us with ~65 million words (~4% of the corpus). For the concrete-only models, we chose words that had a z-value of >0 and removed all other words from the corpus; this left us with ~280 million words (~18% of the corpus). Note that for the concrete-only models, the 61 abstract words were also left in the corpus, so that relationships between them and the concrete words could be obtained.

For all models of each model family, we extracted a 61×61 RDM, which was then correlated with the neural RDM extracted for the ROI; these correlations were Fisher-transformed before statistical analysis. Correlations between all RDMs constructed from the word2vec models can be found in the Supplementary Information (Fig C in S1 Text).

Statistical testing

For the searchlight analyses, to detect spatial clusters in which the neural data were explained by the different representational models, we performed one-sided t-tests against zero across participants, separately for each voxel in the correlation maps. The resulting statistical maps were thresholded at the voxel level at $p_{voxel}<0.001$ (uncorrected) and at the cluster level at $p_{cluster}<0.05$ (family-wise error corrected, as implemented in SPM12). These thresholds were selected based on recommendations for cluster-based thresholding of fMRI results [73]. In the Supplementary Information (Fig F in S1 Text), we show that similar results are reached with an alternative statistical approach based on threshold-free cluster enhancement (TFCE [74]).

For the ROI analyses, correlations between neural RDMs extracted from the IPC ROI and model RDMs were evaluated using one-sided t-tests against zero across participants. Results were corrected for multiple comparisons across the different training corpus sizes using FDR corrections.

Supporting information

S1 Text. Fig A. Neural correlates of visual and auditory word similarity. Visual word similarity, as modelled by the early layers of the an AlexNet DNN (here: layer 3) predicted activations in bilateral posterior visual cortex, including fusiform cortex (866 voxels, peak: -24/-97/-5, t[18] = 7.89). Auditory similarity of the spoken words was modelled by a speech recognition DNN. Early layers of the network (here: layer 2), because of the correlation of word length and speech duration, also predicted activations in bilateral posterior visual cortex (266 voxels, peak: 24/-76/-8, t[18] = 6.51). By contrast, the last layer of the network (layer 7) predicted activations in left middle temporal gyrus (34 voxels, peak: -60/-13/-8, t[18] = 5.61). Brain maps are thresholded at $p_{voxel}<0.001$ (uncorrected) and $p_{cluster}<0.05$ (FWE-corrected). **Fig B. Individual-word effects. a)** Brain-model correlations in left IPC when individual words were deleted from the model RDMs and neural RDMs before performing the analysis. The relatively homogeneous pattern shows that no single word exerted a substantial influence on the correspondence between model and brain. Error margins denote standard errors of the mean. **b)** Brain-model correlations in left IPC when removing a subset of up to 30 words at random from the RDMs. Results across 100 analyses with random subsets removed reveal that the pattern largely holds for smaller subsets of the stimulus space. All data points are means across all participants. **Fig C. Model intercorrelations.** Pairwise correlations between the representational dissimilarity matrices (RDMs) constructed for all word2vec model variants used in the study. **Fig D. Controlling for word frequencies. a)** Whole-brain searchlight analysis when controlling for similarities in word frequency using partial correlations. This analysis yielded results analogous to the original analysis (Fig 1C), with two clusters in right IPC (80 voxels, peak: 45/-

46/58, t[18] = 4.91, and 35 voxels, peak: 36/-73/40, t[18] = 4.79) and one cluster in left IPC (163 voxels, peak: -36/-58/46, t[18] = 6.23). **b)** Region-of-interest analysis in IPC for differently sized training corpora when similarities in word frequencies were controlled for. This analysis revealed essentially identical results to the main analysis (Fig 1E). **c)** Relative frequencies of the 61 abstract words across the differently sized training corpora. Despite some variations across corpus size, the words that were frequent in large corpora were also more frequent in the small corpora. Words are sorted by frequency in the largest (45m) corpus. **Fig E. Direct model comparisons in left IPC.** Differences between in brain-model correlations between the full model (45m sentences and all words included) and the trimmed models. Asterisks indicate significant differences to the full model, FDR-corrected for multiple comparisons. The full model itself is omitted from the plot. Color conventions are the same as in Fig 1E. Error bars denote standard errors. **Fig F. Whole-brain searchlight with TFCE statistics.** Here, we used an alternative statistical test, based on threshold-free cluster enhancement (TFCE; Smith & Nichols, 2009), as implemented in CoSMoMVPA (Oosterhof et al., 2016). Z-scores for TFCE values were obtained by comparing the actual values to values across a null distribution constructed from 10,000 sign permutations. The resulting statistical maps were thresholded at $z>1.96$ ($p<0.05$). As in the main analysis (Fig 1C), two clusters emerged in right parietal cortex (503 voxels, peak: 42/-46/58, z = 2.58, and 88 voxels, peak: -6/-58/58, z = 2.11) and one cluster emerged in left parietal cortex (442 voxels, peak: -27/-73/46, z = 2.89). Although also centered on the IPC, TFCE yielded somewhat more liberal results, with clusters extending more into the superior parietal cortices. No other clusters emerged across the brain. **Fig G. Searchlight results on cross-sectional images.** Coronal brain slices are overlaid with regions that show significant correlations between neural representations and the full word2vec model (as in Fig 1C). **Fig H. Unthresholded searchlight results.** Coronal brain slices are overlaid with unthresholded t-maps comparing brain-model correlations to zero for the full word2vec model (as in Fig 1C). Slices are spaced continuously between z = -34 and z = 54. Negative t-values are truncated. **Table A. Abstract words.** The 61 abstract German words used in the experiment and their English translations. **Table B. Background contexts.** The 10 background contexts used in the experiment and their English translations. The gray text on the bottom was always used during the practice run.
(PDF)

Acknowledgments

Thanks to Raphael Leuner and Kshitij Dwivedi for help with the computational analyses.

Author Contributions

Conceptualization: Daniel Kaiser, Radoslaw M. Cichy.

Data curation: Daniel Kaiser.

Formal analysis: Daniel Kaiser, Arthur M. Jacobs.

Funding acquisition: Daniel Kaiser, Radoslaw M. Cichy.

Investigation: Daniel Kaiser.

Methodology: Daniel Kaiser, Arthur M. Jacobs, Radoslaw M. Cichy.

Project administration: Daniel Kaiser, Radoslaw M. Cichy.

Resources: Daniel Kaiser, Arthur M. Jacobs.

Software: Daniel Kaiser, Arthur M. Jacobs.

Supervision: Daniel Kaiser, Radoslaw M. Cichy.

Validation: Daniel Kaiser, Radoslaw M. Cichy.

Visualization: Daniel Kaiser.

Writing – original draft: Daniel Kaiser.

Writing – review & editing: Daniel Kaiser, Arthur M. Jacobs, Radoslaw M. Cichy.

References

1. Binder JR, Desai RH, Graves WW, Conant LL. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex. 2009; 19: 2767–2797. https://doi.org/10.1093/cercor/bhp055 PMID: 19329570
2. Deniz F, Nunez-Elizalde AO, Huth AG, Gallant JL. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. J Neurosci. 2019; 39: 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019 PMID: 31427396
3. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Human natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 2016; 532: 453–458. https://doi.org/10.1038/nature17637 PMID: 27121839
4. Just MA, Cherkassky VL, Aryal S, Mitchell TM. A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS One 2010; 5: e8622. https://doi.org/10.1371/journal.pone.0008622 PMID: 20084104
5. Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, et al. Predicting human brain activity associated with the meanings of nouns. Science. 2008; 320: 1191–1195. https://doi.org/10.1126/science.1152876 PMID: 18511683
6. Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, et al. Toward a universal decoder of linguistic meaning from brain activation. Nat Commun 2018; 9: 1–13. https://doi.org/10.1038/s41467-017-02088-w PMID: 29317637
7. Anderson AJ, Kiela D, Clark S, Poesio M. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. Trans Assoc Comput Linguist. 2017; 5: 17–30.
8. Anderson AJ, Murphy B, Poesio M. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. J Cogn Neurosci. 2014; 26: 658–681. https://doi.org/10.1162/jocn_a_00508 PMID: 24168217
9. Vargas R, Just MA. Neural representations of abstract concepts: identifying underlying neurosemantic dimensions. Cereb Cortex. 2020; 30: 2157–2166. https://doi.org/10.1093/cercor/bhz229 PMID: 31665238
10. Wang X, Wu W, Ling Z, Xu Y, Fang Y, Wang X, et al. Organizational principles of abstract words in the human brain. Cereb Cortex. 2018; 28: 4305–4318. https://doi.org/10.1093/cercor/bhx283 PMID: 29186345
11. Borghi AM, Binkofski F, Castelfranchi C, Cimatti F, Scorolli C, Tummolini L. The challenge of abstract concepts. Psychol Bull- 2017; 143: 263–292. https://doi.org/10.1037/bul0000089 PMID: 28095000
12. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis–connecting the branches of systems neuroscience. Front Syst Neurosci. 2008; 2: 4. https://doi.org/10.3389/neuro.06.004.2008 PMID: 19104670
13. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781 [preprint]. 2013. Available from: https://arxiv.org/abs/1301.3781
14. Faaß G, Eckhart K. SdeWaC–a corpus of parsable sentences from the web. In: Gurevych I, Biemann C, Zesch T, editors. Language processing and knowledge in the web. Lecture notes in computer science, vol. 8105. Berlin: Springer; 2013.
15. Huebner PA, Willits JA. Structured semantic knowledge can emerge automatically from predicting words sequences in child-directed speech. Front Psychol. 2018; 9: 133. https://doi.org/10.3389/fpsyg.2018.00133 PMID: 29520243
16. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2019; 571: 95–98. https://doi.org/10.1038/s41586-019-1335-8 PMID: 31270483

17. Utsumi A. Exploring what is encoded in distributional word vectors: a neurobiologically motivated analysis. Cogn Sci. 2020; 44: e1284. https://doi.org/10.1111/cogs.12844 PMID: 32458523

18. Kousta ST, Vigliocco G, Vinson DP, Andrews M, Del Campo E. The representation of abstract words: why emotion matters. J Exp Psychol Gen. 2011; 140: 14. https://doi.org/10.1037/a0021446 PMID: 21171803

19. Vigliocco G, Kousta ST, Della Rosa PA, Vinson DP, Tettamanti M, Devlin JT, et al. The neural representation of abstract words: the role of emotion. Cereb Cortex. 2011; 24: 1767–1777.

20. Freud E, Culham JC, Plaut DC, Behrmann M. The large-scale organization of shape processing in the ventral and dorsal pathways. eLife. 2018; 6: e27576.

21. Hartwigsen G, Baumgaertner A, Price CJ, Koehnke M, Ulmer S, Siebner HR. Phonological decisions require both the left and right supramarginal gyri. Proc Natl Acad Sci USA. 2010; 107: 16494–16499. https://doi.org/10.1073/pnas.1008121107 PMID: 20807747

22. Kiefer M, Pulvermüller F. Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. Cortex. 2012; 48: 805–825. https://doi.org/10.1016/j.cortex.2011.04.006 PMID: 21621764

23. Lakoff G. The neural theory of metaphor. In: Gibbs WR Jr, editor. The Cambridge handbook of metaphor and thought. Cambridge: Cambridge University Press; 2008.

24. Braga RM, DiNicola LM, Becker HC, Buckner RL. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. J Neurophysiol. 2020; 124: 1415–1448. https://doi.org/10.1152/jn.00753.2019 PMID: 32965153

25. Binder JR, Desai RH. The neurobiology of semantic memory. Trends Cogn Sci. 2011; 15: 527–536. https://doi.org/10.1016/j.tics.2011.10.001 PMID: 22001867

26. Humphreys GF, Hoffman P, Visser M, Binney RJ, Lambon Ralph MA. Establishing task- and modality-dependent dissociations between the semantic and default mode networks. Proc Natl Acad Sci USA. 2015; 112: 7857–7862. https://doi.org/10.1073/pnas.1422760112 PMID: 26056304

27. Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT. The neural and computational bases of semantic cognition. Nat Rev Neurosci. 2017; 18: 42–55. https://doi.org/10.1038/nrn.2016.150 PMID: 27881854

28. David CP, Yee E. Features, labels, space, and time: factors supporting taxonomic relationships in the anterior temporal lobe and thematic relationships in the angular gyrus. Lang Cogn Neurosci. 2019; 34: 1347–1357.

29. Graessner A, Zaccarella E, Hartwigsen G. Differential contributions of left-hemispheric language regions to basic semantic composition. Brain Struct Funct. 2021; 226: 501–518. https://doi.org/10.1007/s00429-020-02196-2 PMID: 33515279

30. Price AR, Bonner MF, Peelle JE, Grossman M. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. J Neurosci. 2015; 35: 3276–3284. https://doi.org/10.1523/JNEUROSCI.3446-14.2015 PMID: 25698762

31. Pylkkänen L. The neural basis of combinatory syntax and semantics. Science 2019; 355: 62–66. https://doi.org/10.1126/science.aax0050 PMID: 31604303

32. Pylkkänen L. Neural basis of basic composition: what we have learned from the red-boat studies and their extensions. Phil Trans R Soc Lond B Biol Sci. 2020; 375: 20190299. https://doi.org/10.1098/rstb.2019.0299 PMID: 31840587

33. Hoffman P, Binney RJ, Lambon Ralph MA. Differing contributions of inferior prefrontal and anterior temporal cortex to concrete and abstract conceptual knowledge. Cortex. 2015; 63: 250–266. https://doi.org/10.1016/j.cortex.2014.09.001 PMID: 25303272

34. Binder JR, Westbury CF, McKiernan KA, Possing ET, Medler DA. Distinct brain systems for processing concrete and abstract concepts. J Cogn Neurosci. 2005; 17: 905–917. https://doi.org/10.1162/0898929054021102 PMID: 16021798

35. Wang J, Conder JA, Blitzer DN, Shinkareva SV. Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. Hum Brain Mapp. 2010; 31: 1459–1468. https://doi.org/10.1002/hbm.20950 PMID: 20108224

36. Harpaintner M, Sim E-J, Trumpp NM, Ulrich M, Kiefer M. The grounding of abstract concepts in the motor and visual system: an fMRI study. Cortex. 2020; 124: 1–22. https://doi.org/10.1016/j.cortex.2019.10.014 PMID: 31821905

37. Tang J, LeBel A, Huth AG. Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. bioRxiv [preprint]. 2021. Available from: https://doi.org/10.1101/2021.05.19.444701

38. Dreyer FR, Pulvermüller F. Abstract semantics in the motor system?–An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. Cortex. 2018; 100: 52–70. https://doi.org/10.1016/j.cortex.2017.10.021 PMID: 29455946

39. Firth JR. A synopsis of linguistic theory, 1930–1950. In: Studies in Linguistic Analysis. Oxford: Blackwell; 1957. https://doi.org/10.1111/j.1471-0528.1957.tb02658.x PMID: 13449662

40. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev. 1997; 104: 211–240.

41. Harris ZS. Distributional structure. Word. 1954; 10: 146–162.

42. Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5185–5198.

43. Vigliocco G, Ponari M, Norbury C. Learning and processing abstract words and concepts: insights from typical and atypical development. Top Cogn Sci. 2018; 10: 533–549. https://doi.org/10.1111/tops.12347 PMID: 29785838

44. Cichy RM, Kaiser D. Deep neural networks as scientific models. Trends Cogn Sci. 2019; 23: 305–317. https://doi.org/10.1016/j.tics.2019.01.009 PMID: 30795896

45. Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, et al. The neural architecture of language: integrative modeling converges on predictive processing. Proc Natl Acad Sci USA. 2021; 118: e2105646118. https://doi.org/10.1073/pnas.2105646118 PMID: 34737231

46. Anderson AJ, Kiela D, Binder JR, Fernandino L, Humphries CJ, Conant LL, et al. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. J Neurosci. 2021; 41: 4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021 PMID: 33753548

47. Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, et al. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. bioRxiv [preprint]. 2021. Available from: https://doi.org/10.1101/2020.12.02.403477.

48. Jain S, Huth AG. Incorporating context into language encoding models for fMRI. Advances in Neural Information Processing Systems, 2018: 31.

49. Lopopolo A, Schoffelen JM, van den Bosch A, Willems RM. Words in context: tracking context-processing during language comprehension using computational language models and MEG. bioRxiv [preprint]. 2020. Available from: https://doi.org/10.1101/2020.06.19.161190.

50. Bonner MF, Epstein RA. Object representations in the human brain reflect the co-occurrence statistics of vision and language. Nat Commun. 2021; 12: 4081. https://doi.org/10.1038/s41467-021-24368-2 PMID: 34215754

51. Hayes TR, Henderson JM. Looking for semantic similarity: what a vector-space model of semantics can tell us about attention in real-world scenes. Psychol Sci, forthcoming 2022.

52. van Paridon J, Liu Q, Lupyan G. How do blind people know that blue is cold? Distributional semantics encode color-adjective associations. PsychaRxiv [preprint]. 2021. Available from: https://doi.org/10.31234/osf.io/vyxpq

53. Greiner B. Subject pool recruitment procedures: organizing experiments with ORSEE. JESA. 2015; 1: 114–125.

54. Brainard DH. The psychophysics toolbox. Spat Vis. 1997; 10: 433–436. PMID: 9176952

55. Martin CB, Douglas D, Newsome RN, Man LLY, Barense MD. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. eLife. 2018; 7: e31873. https://doi.org/10.7554/eLife.31873 PMID: 29393853

56. Wang X, Xu Y, Wang Y, Zeng Y, Zhang J, Ling Z, et al. Representational similarity analysis reveals task-dependent semantic influence on the visual word form area. Sci Rep. 2018; 8: 3047. https://doi.org/10.1038/s41598-018-21062-0 PMID: 29445098

57. van Gerven MAJ. A primer on encoding models in sensory neuroscience. J Math Psychol. 2017; 76: 172–183.

58. Diedrichsen J, Kriegeskorte N. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. Plos Comput Biol. 2017; 13: e1005508. https://doi.org/10.1371/journal.pcbi.1005508 PMID: 28437426

59. Kriegeskorte N, Douglas PK. Interpreting encoding and decoding models. Curr Opin Neurobiol. 2019; 55: 167–179. https://doi.org/10.1016/j.conb.2019.04.002 PMID: 31039527

60. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proc Natl Acad Sci USA. 2006; 103: 3863–3868. https://doi.org/10.1073/pnas.0600244103 PMID: 16537458

61. Oosterhof NN, Connolly AC, Haxby JV. CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. Front Neuroinform. 2016; 10: 20. https://doi.org/10.3389/fninf.2016.00020 PMID: 27378902

62. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001; 293: 2425–2430. https://doi.org/10.1126/science.1063736 PMID: 11577229

63. Walther A., Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage. 2016; 137: 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012 PMID: 26707889

64. Russell JA, Weiss A, Mendelsohn GA. Affect grid: a single-item scale of pleasure and arousal. J Personal Soc Psychol. 1989; 57: 493.

65. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012: 1097–1105.

66. Vedaldi A, Lenc K. MatConvNet–convolutional neural networks for Matlab. ACM International Conference on Multimedia. 2015.

67. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep. 2016; 6: 27755. https://doi.org/10.1038/srep27755 PMID: 27282108

68. Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron. 2018; 98: 630–644. https://doi.org/10.1016/j.neuron.2018.03.044 PMID: 29681533

69. Ambrus GG, Kaiser D, Cichy RM, Kovács G. The neural dynamics of familiar face recognition. Cereb Cortex. 2019; 29: 4775–4784. https://doi.org/10.1093/cercor/bhz010 PMID: 30753332

70. Cichy RM, Kriegeskorte N, Jozwik KM, van den Bosch JJF, Charest I. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. Neuroimage. 2019; 194: 12–24. https://doi.org/10.1016/j.neuroimage.2019.03.031 PMID: 30894333

71. Kaiser D, Nyga K. Tracking cortical representations of facial attractiveness using time-resolved representational similarity analysis. Sci Rep. 2020; 10: 16852. https://doi.org/10.1038/s41598-020-74009-9 PMID: 33033356

72. Köper M, im Walde SS. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. Proceedings of the 1st workshop on sense, concept and entity representations and their applications. 2017: 24–30.

73. Woo C-W, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage. 2014; 91: 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058 PMID: 24412399

74. Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localization in cluster inference. Neuroimage. 2009; 44: 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061 PMID: 18501637