Distributed Semantic Representations in the Human Brain: Evidence from Natural Language Processing Techniques

Authors: Jiang Jiahao, Zhao Guoyu, Yingbo Ma, Ding Guosheng, Liu Lanfang, Liu Lanfang

Date: 2023-01-18T00:00:00+00:00

Abstract

How the human brain represents semantic information has long been a central question in cognitive neuroscience. Traditional research has primarily employed experimental methods that artificially manipulate stimulus properties or task demands to localize brain regions involved in semantic representation. While these approaches have yielded numerous achievements, they still suffer from limitations such as the difficulty in quantifying semantic information and contextual effects in detail. Grounded in the distributed hypothesis of semantics, natural language processing (NLP) techniques transform discrete, objectively unquantifiable semantic information into a unified, computable vector representation, substantially enhancing the precision of semantic characterization and providing effective tools for quantifying contextual and syntactic information. By utilizing NLP techniques to extract stimulus semantic information and establishing mapping relationships between semantic vectors and brain activity patterns through representational similarity analysis or linear regression, researchers have discovered that neural structures representing semantic information are widely distributed across multiple brain regions, including the temporal lobe, frontal lobe, and occipital lobe. Future research may incorporate more sophisticated semantic representation methods such as knowledge graphs and multimodal fusion models, employ language models to evaluate language abilities in special populations, or leverage cognitive neuroscience experiments to enhance the interpretability of deep language models.

Full Text

Distributed Representation of Semantics in the Human Brain: Evidence from Natural Language Processing Techniques

JIANG Jiahao[1], ZHAO Guoyu[2], MA Yingbo[1], DING Guosheng[3], LIU Lanfang[2],[4]

([1] Faculty of Psychology, Beijing Normal University, Beijing 100875, China)
([2] Department of Psychology, School of Arts and Sciences, Beijing Normal University at Zhuhai, Zhuhai 519087, China)
([3] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University & IDG/McGovern Institute for Brain Research, Beijing 100875, China)
([4] Center for Cognition and Neuroergonomics, Beijing Normal University at Zhuhai, Zhuhai 519087, China)

Abstract: How the human brain represents semantic information has long been a central question in cognitive neuroscience. Traditional research has primarily employed experimental methods that manipulate stimulus properties or task demands to localize semantic brain regions. While these approaches have yielded numerous insights, they face challenges in quantifying semantic information in detail and accounting for contextual effects. Based on the distributed hypothesis of semantics, natural language processing (NLP) technologies transform discrete, difficult-to-quantify semantic information into uniform, computable vector representations, dramatically improving the precision of semantic characterization and providing effective tools for quantifying contextual and syntactic information. By extracting stimulus semantic information using NLP techniques and establishing mapping relationships between semantic vectors and brain activity patterns through representational similarity analysis or linear regression, researchers have discovered that neural structures representing semantic information are widely distributed across multiple brain regions including the temporal, frontal, and occipital lobes. Future research may introduce more complex semantic representation methods such as knowledge graphs and multimodal fusion models, apply language models to assess language abilities in special populations, or utilize cognitive neuroscience experiments to improve the interpretability of deep language models.

Keywords: semantic representation, brain, natural language processing, language model

Language, as an abstract symbolic system, represents humanity' s most important tool for expressing meaning and exchanging information. Through the combination of a finite set of linguistic units, people can comprehend and express infinite information encompassing knowledge, beliefs, intentions, emotions, and more. Revealing how the human brain stores, accesses, and retrieves semantics has remained one of the core questions in cognitive neuroscience. To investigate

the neural basis of semantic representation and processing, researchers have traditionally manipulated stimulus attributes or task requirements, comparing brain activation patterns across different conditions. For example, studies have contrasted brain activation differences between real words and pseudowords in lexical decision tasks [**?**, **?**], or compared brain activity during semantic versus phonological judgment tasks for identical linguistic stimuli [**?**, **?**]. While this paradigm of strict experimental control and condition comparison has yielded important findings, it suffers from several limitations when exploring the neural representation and processing of semantics.

First, characterizing semantic features relies on manual ratings with coarse granularity. Daily communication contexts are complex and variable, yet people only need to master a small vocabulary to meet conversational needs—for instance, just 590 Chinese characters cover 80% of daily usage [**?**, **?**]. A limited set of characters can combine to express infinite meanings because people construct rich mental representations for each lexical item, with subtle differences across multiple dimensions. Based on psychological experiments or linguistic classifications, current research on semantic relationships mostly remains at a coarse-grained level, such as distinguishing nouns from verbs or living versus non-living concepts. To refine semantic representation, recent researchers have measured conceptual words along psychological dimensions, such as using 12 dimensions including time, space, quantity, and arousal to characterize abstract conceptual words [**?**, **?**], or employing 65 experiential dimensions encompassing sensory, motor, temporal, spatial, and social cognitive components to represent concepts [**?**, **?**]. While this psychological-dimension approach can capture both concepts themselves and inter-concept relationships with high interpretability, it still has limitations. For instance, dimension selection is subjectively determined by researchers, whose reasonableness and completeness require validation. Moreover, quantifying word meanings primarily through subjective participant judgments renders results susceptible to individual knowledge and experience. Finally, the rating method is time-consuming and labor-intensive, difficult to generalize to all vocabulary, unable to comprehensively cover multiple meanings of words across different contexts, and the varying word lists and dimensions selected by different researchers increase the difficulty of comparing and integrating findings.

Second, contextual effects are difficult to quantify. In language systems worldwide, most characters or words can refer to multiple meanings—for example, over 80% of English words exhibit polysemy [**?**, **?**]. In real-world situations, the meaning of a linguistic symbol activated in an individual depends heavily on context; in other words, the representation and retrieval of linguistic meaning is dynamic and context-dependent [**?**, **?**]. For instance, mentioning "air conditioner" in summer versus winter tends to evoke opposite functions. However, due to the inherent complexity of context, it is difficult to objectively measure contextual effects through experimental design. Consequently, most current studies use highly controlled materials such as isolated linguistic stimuli or sentences with scrambled syntax or semantics, which still differ substantially from

everyday language use. Answering questions about how the brain represents and processes context, and how semantic representations are dynamically influenced by contextual information, remains a significant challenge.

Third, discourse-level thematic information is difficult to quantify. Discourse (e.g., news reports, stories) consists of words and sentences connected through complex relationships, with semantic associations between different parts that express complete, coherent meaning (themes). To investigate the processing and representation of discourse-level semantic information, psychology researchers typically compare intact discourse with materials scrambled at different levels (words, sentences, or paragraphs) [**?**, **?**, **?**, **?**]. However, scrambled materials present greater complexity and difficulty (potentially eliciting stronger brain activation), and people tend to attempt reorganizing and integrating scrambled materials to achieve semantic coherence. Therefore, subtraction methods may not accurately detect processing specific to discourse-level semantics. Additionally, this experimental approach cannot measure the semantic structural relationships within discourse or semantic distances between different discourses.

Given the limitations of traditional psychological experimental methods, an increasing number of psychology researchers have recently introduced natural language processing (NLP) techniques from artificial intelligence, particularly neural network-based and deep learning language models, to measure the semantics and semantic relationships of experimental stimuli. Combining NLP models with brain imaging experimental data is becoming an important trend in neurolinguistics. Some domestic and international researchers have recently summarized and prospectively reviewed the application of computational linguistic methods in cognitive linguistics and brain science. For example, Wang et al. (2022b) summarized the application of emerging computational linguistic methods to questions concerning the units and dimensions of linguistic information, brain network localization of different types of linguistic information, temporal dynamics and control of linguistic information processing, and the neural encoding forms and computational mechanisms of linguistic information, covering multiple aspects including phonology, semantics, and syntactic structure. In another article [**?**, **?**], the authors systematically discussed the research questions, methods, and limitations of cognitive linguistics and computational linguistics from a macro perspective, offering profound insights into how these two fields might integrate. Other researchers have conducted in-depth comparisons between modern distributed semantic computational models and two traditional semantic models in cognitive psychology (feature-based semantic models and connection network-based semantic models) regarding knowledge representation, learning mechanisms, and semantic disambiguation, and explored pathways for combining modern semantic computational models with these two traditional approaches [**?**, **?**].

While the aforementioned studies have broadly overviewed the extensive applications of computational linguistic methods in language cognition, they have not systematically summarized or elaborated on specific issues. This review

focuses on one of the core issues in language cognition and brain science—the representation of semantic information in the human brain—and summarizes and prospects the application of NLP models to this question. This review will first introduce the principles and techniques of semantic representation in NLP models and two methods for combining language models with brain imaging data. On this basis, it will systematically elaborate on the application of NLP techniques in research on human brain semantic representation, including word-level semantics, sentence-level (and contextual) semantics, and discourse-level semantics, and compare these with the limitations of traditional psychological methods for measuring semantics. Finally, it will discuss potential pitfalls, challenges, and future directions in applying NLP language models to investigate human brain semantic representation.

2. Algorithmic Principles and Advances in NLP Semantic Representation

How to enable computers to automatically capture semantics from text is a core question in computational linguistics. Early researchers proposed logic rule-based approaches to model natural language [**?**, **?**, **?**], hoping computers could understand word meanings based on syntax, word order, and collocation rules like humans do. Although this approach achieved high precision, it heavily relied on manually compiled linguistic grammars, making it unsuitable for processing large-scale real-world texts (especially in the Internet era with increasingly frequent new word usages and meanings), and rules varied across different languages. Later, due to many problems with rule-based representation, the statistical school proposed the vector space model of semantics [**?**, **?**] based on the distributed semantic hypothesis that "words appearing in similar contexts have similar meanings" [**?**, **?**], which has become the mainstream guiding ideology in the NLP field for over a decade—distributed representation. This idea maps discrete symbols (local representation) like words into a dense vector space, using a relatively low-dimensional vector (e.g., 300 dimensions) instead of a sparse one-hot vector of hundreds of thousands of dimensions [**?**, **?**]. For example, the local representation of colors is "red, orange, yellow, gray, Chinese red ⋯" ([1,0,0,0,0], [0,1,0,0,0], [0,0,1,0,0], [0,0,0,1,0], [0,0,0,0,1]), whereas distributed representation can unify all colors into a three-dimensional RGB vector (e.g., gray can be represented as [125, 125, 125]), greatly reducing vector dimensions. In distributed representation, semantic information is implicitly encoded in various dimensions of word vectors, and semantic relationships between words are primarily reflected by their spatial positional relationships: the closer two word vectors are, the higher their semantic similarity.

Regarding the construction of semantic spaces and acquisition of word vectors, there are currently two main approaches. One is statistical-based semantic representation methods, which primarily rely on corpus statistics of co-occurrence relationships between "word-word" or "word-document" pairs, including algorithms such as Latent Semantic Analysis (LSA) [**?**, **?**, **?**], Non-negative Matrix

Factorization (NMF) [?, ?], and N-gram based on Markov assumptions [?, ?]. Taking LSA as an example, this method establishes a "word-document" co-occurrence matrix through statistical analysis of text corpora, where the matrix is the number of documents, then performs singular value decomposition on the co-occurrence matrix × to build a latent semantic space and achieve dimensionality reduction (the formula represents the dimensionality of the latent semantic space). In the matrix decomposition × = × × × , each row represents the latent semantic representation of a word (i.e., word vector), each column in the matrix represents the latent semantic representation of a document, and the singular values in the matrix reflect the importance of each latent semantic dimension. In this way, both words and documents are mapped into a unified latent semantic space with condensed information, enabling representation of both word semantics and document semantics. Statistical-based semantic representation methods can effectively cluster semantically similar words or documents and have achieved good performance in tasks such as semantic similarity analysis, word/document clustering, and information extraction [?, ?, ?]. However, this method also has obvious limitations: for example, the distribution of word (or document) vectors may not conform to the normal distribution required by probabilistic model assumptions; matrix decomposition has high computational complexity, and when new documents are added, the model must be retrained; it fails to adequately consider word order information in sentences; and it cannot resolve polysemy.

Different from statistical-based methods, the other approach uses neural networks to learn semantic representation by adjusting model parameters based on differences between predicted and actual values (for other classification standards of semantic modeling methods, please refer to [?, ?]). Artificial Neural Networks (ANN, hereafter referred to as neural networks) are mathematical models constructed by simulating the complex information processing mechanisms of the human nervous system [?, ?]. Neural networks consist of neurons (nodes) interconnected by edges, primarily comprising an input layer, hidden layers, and an output layer in sequence. The input layer mainly receives and activates signals (e.g., extracting word vectors corresponding to words, analogous to external stimuli eliciting electrophysiological activity in primary sensory areas); hidden layers are the core of neural networks, performing complex processes such as signal processing, integration, and abstraction (analogous to interneurons, association cortices, and high-level decision-making cortices in the brain); the output layer receives signals processed by hidden layers and produces final responses according to task demands (e.g., classifying words by emotion, analogous to the brain' s articulatory and motor cortices). Similar to the characteristics of neuronal action potentials in the brain, neurons in hidden layers of artificial neural networks receive signals from multiple upstream neurons (analogous to dendrites), perform weighted summation according to different weights (analogous to the cell body), and then decide whether to transmit signals downstream and with what intensity based on whether the aggregated signal exceeds an activation threshold (generally completed through non-linear activation func-

tions like sigmoid or ReLU), with subsequent hidden layers working similarly. Notably, the information weights between each neuron in hidden layers and various upstream neurons differ, and these parameters are continuously adjusted through backpropagation algorithms based on errors between network outputs and actual values. Through multiple training iterations that progressively reduce the gap between predicted and actual values, neural networks establish mapping relationships between original input signals and target outputs, with final learning outcomes reflected in the parameters of individual neurons.

Regarding word vector representation, neural networks typically use large-scale corpora to train network weights, inputting sentence materials to learn relationships between words and contextual environments. Taking the classic Continuous Bag-of-Words (CBOW) model in Word2Vec as an example [**?**, **?**], this model is designed based on the distributed hypothesis (words with similar contexts have similar meanings), predicting the target word in the middle given a total of context words before and after it. The input layer consists of one-hot encoded vectors of words, extracting word vectors through weight matrices between the input and hidden layers, then performing dot product multiplication with the weight matrix between the hidden and output layers, followed by softmax normalization to obtain probabilities of each word in the vocabulary appearing, selecting the word with the highest probability as the prediction result (see Figure 1 [Figure 1: see original paper]). By calculating differences between predicted and actual word vectors and adjusting parameters through backpropagation, the weights between the input and hidden layers (i.e., word vectors) are continuously updated. Additionally, Word2Vec can be trained using the skip-gram model, which predicts the context (a total of words before and after) given a target word. Word2Vec-derived word vectors align well with the distributed hypothesis, produce reasonable clustering results, and can effectively reflect semantic similarity [**?**, **?**, **?**]. For example, calculating the vector ( ) = − + ( ) yields the highest cosine similarity with word vectors of related words such as.

Following the proposal of the Word2Vec model, the NLP field witnessed a surge in word vector computation and optimization research. Subsequent researchers designed a series of neural network language models with more complex architectures that consider contextual information when computing word vectors, better aligning with the human brain' s cognitive mode of context integration. Newly developed neural network models can also model sentence and discourse semantics, with representative models including: Recursive Neural Networks (RecNN) that can capture sentence structural information [**?**, **?**]; Recurrent Neural Networks (RNN) [**?**, **?**, **?**] and their optimized version Long Short-Term Memory networks (LSTM) [**?**, **?**, **?**], which treat sentences as sequential time series and integrate upstream (downstream) context information into current word vector representations [**?**, **?**]; and Convolutional Neural Networks (CNN), which extract multi-level semantic information and possess more efficient parallel computing capabilities [**?**, **?**, **?**]. In addition to words, neural network-based algorithms can also represent paragraph or discourse semantics. For example, Doc2Vec adds a

paragraph vector that is shared within paragraphs but independent across paragraphs for training on top of the Word2Vec model, thereby obtaining vectorized semantic representations of paragraphs [**?**, **?**]. Other approaches include hierarchical feature extraction, such as first computing semantic representations of each sentence within a paragraph to obtain sentence vectors, then using sentence vectors as input units to obtain paragraph vectors.

Later, Google proposed the Transformer architecture [**?**, **?**], which addressed the limitations of long-distance dependencies and inefficient serial training in RNNs and their variants, becoming the mainstream network backbone for recent NLP models. The Transformer architecture consists of encoders and decoders, each containing multi-head self-attention layers and fully connected layers. The self-attention layer integrates contextual information by calculating and weighted summing similarities between target words and context words, followed by feature extraction through fully connected layers. The self-attention mechanism in Transformer replaces the serial memory units in RNN structures, enabling high-speed parallelization of computations, and the architecture enhances text feature extraction and abstraction effects through stacking multiple encoders and decoders. Representative language models based on the Transformer architecture include BERT (Bidirectional Encoder Representation from Transformers) [**?**, **?**] and GPT (Generative Pre-trained Transformer) [**?**, **?**, **?**], which have achieved significant improvements in many natural language processing tasks. Deep neural network-based semantic modeling methods have enormous parameters (e.g., the BERT-large model has 300 million parameters to train, while GPT-3 has as many as 175 billion parameters), requiring high demands on corpus data volume and computer performance. Therefore, pre-training has become the mainstream usage method for large-scale language models, training the model extensively on a language task (e.g., cloze tests) to obtain model parameters, with research teams using this set of model parameters as a foundation for downstream tasks. Pre-trained models reduce the technical and time costs of model training for research teams and enhance the comparability and reproducibility of language cognition research.

Compared with traditional statistical-based semantic representation methods, neural network models can capture richer text features with stronger generalizability, demonstrating superior performance in various complex language tasks such as cloze tests, sentiment analysis, abstract generation, and translation [**?**, **?**, **?**]. Moreover, large-scale pre-trained models (e.g., BERT) encode various types of linguistic information within their parameters, allowing researchers to fine-tune pre-trained models according to their needs, thereby obtaining better model performance for specialized tasks with lower resource consumption. With continuous improvements in computing power, these advantages and performance have enabled neural network models to gradually replace traditional statistical-based text representation methods as one of the core technologies in the NLP field. For more detailed introductions to text representation methods in NLP, please refer to [**?**, **?**].

3.1 Methods for Combining NLP Language Models with Brain Imaging Data

NLP language models provide effective tools for objectively measuring and computing text semantics. Using these tools, neurolinguistics researchers can further analyze the extent to which semantic information explains variations in brain activity patterns, thereby inferring which brain regions participate in semantic representation and processing. Notably, word vectors derived from NLP language models and brain activity data come from different models and modalities, with completely different data dimensions and meanings. For example, vectors from BERT' s output layer are 768-dimensional (BERT-base) or 1024-dimensional (BERT-large), with unclear meanings for each dimension. Brain activity data dimensions vary according to selected brain region sizes, ranging from one dimension (voxel level), hundreds (ROI level), thousands (network level), to tens of thousands (whole-brain level). How to effectively model these two types of multivariate data with different dimensions is a challenging problem, with two commonly used methods currently: Representational Similarity Analysis (RSA) and linear regression.

RSA establishes relationships between two types of data by analyzing the shared structure between semantic similarity matrices and brain activity similarity matrices [**?**, **?**]. When conducting RSA, it is first necessary to extract representations of various stimuli (e.g., words) from both the human brain and NLP language models, where brain representations can be expressed as activity intensity data of a set of voxels elicited by a given word, and NLP model representations can be expressed as word vectors for that word from Word2Vec (or other models). Then, the representational similarity within the human brain and language models for different stimuli is calculated separately (using metrics such as correlation coefficients, Euclidean distance, or Mahalanobis distance), thereby constructing Representation Dissimilarity Matrices (RDM). RDMs reflect differences in how the same model represents different stimuli. By calculating Spearman similarity between two RDMs, the resulting correlation coefficient reflects the degree of similarity in internal representations of the same set of stimuli between the human brain and language models (see Figure 2 [Figure 2: see original paper]).

Linear regression is another method for associating different types of high-dimensional data. Its basic idea is to find a set of parameters that fit the relationship between two datasets, thereby "predicting" brain responses based on stimulus features or model output vectors (encoding), or "predicting" what content participants are currently processing based on brain activity patterns (decoding). Among various linear regression methods, ridge regression is the most commonly used, as it can address problems such as overfitting and multicollinearity. Many recent studies have found that NLP model vectors can establish mapping relationships with brain activity through ridge regression for the same linguistic information [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. If the model and human brain share similar representational information, there will be significant

correlation between ridge regression predictions and actual values.

Both RSA and ridge regression can compare relationships between different models and brain representations, but they differ in principle and function [?, ?]. RSA measures the similarity degree between NLP model response patterns and patterns of a set of voxels (or brain regions) in the brain, while ridge regression aims to establish regression relationships between features (or model vectors) and activity of individual voxels (or brain regions). The RSA method does not require parameter fitting, thus having lower computational costs and relatively low data requirements. However, this method treats all features as a whole and cannot estimate the contribution of individual features to brain activity. Ridge regression can obtain weight values of individual features on brain activity, thereby predicting activation patterns for new stimuli based on their features, which is more common in tasks using continuous natural stimuli. However, this method requires estimating many free parameters and often necessitates grid search for penalty coefficients, resulting in higher computational costs and greater data requirements. Addressing the respective advantages and disadvantages of RSA and ridge regression, Anderson et al. (2016) proposed the representational similarity encoding method. Based on the idea that "similar stimuli evoke similar brain activities," this method first calculates feature similarities between the target to be predicted and all known targets, then uses similarity metrics as weights to perform weighted averaging of brain activity values evoked by known targets, thereby obtaining predicted brain activity values for the target. This method utilizes similarity information between stimuli for prediction, avoiding model parameter estimation, enabling fast computation, and producing parameters (similarities) in the regression model with strong interpretability and significant application value [?, ?, ?].

Notably, interpretation of correlation coefficients between predicted and actual values in RSA or ridge regression requires caution. Significant correlation coefficients only indicate that model and brain representations share similar information and cannot directly infer that their underlying mechanisms are identical, especially when correlation coefficients are low [?, ?, ?].

3.2.1 Word-Level Semantic Representation

As the carrier of thought, which brain regions process the meaningful information contained in language and how they process it has always been a concern in cognitive neuroscience. Early semantic representation research primarily investigated which brain regions process words or concepts by comparing brain activation differences when participants received different stimuli or performed different tasks, such as contrasts between real and pseudowords [?, ?], word categories [?, ?, ?], parts of speech [?, ?, ?], and semantic versus phonological tasks [?, ?]. The condition-contrast paradigm and activation analysis have yielded many important findings, but characterization of semantic information remains at a coarse-grained level and is difficult to quantify. NLP techniques enable researchers to quantitatively measure semantic information of materials and

explore associations between semantic information and brain representation.

In early work, Mitchell et al. (2008) selected noun stimuli and used their co-occurrence frequencies with 25 representative verbs as semantic vector representations to predict brain activity during noun processing through linear regression. Results showed that bilateral occipital lobes, parietal lobes, and middle frontal gyri could all distinguish words, suggesting that brain representation of concrete nouns is partly based on sensory-motor features, with effects in the occipital lobe possibly due to participants' associations with action scenes related to the nouns. This study pioneered the combination of NLP and brain imaging techniques, providing a new approach beyond condition-contrast paradigms for semantic brain representation research. Recent researchers have begun applying NLP methods to semantic analysis of natural continuous linguistic materials (e.g., stories or movie audio) [**?**, **?**, **?**]. Compared with traditional laboratory methods (artificially compiling or selecting small amounts of specific linguistic stimuli), these natural continuous materials contain larger and more diverse vocabularies, potentially yielding results that better reflect real human brain semantic representation. For example, in Huth et al. (2016), participants listened to a two-hour story while undergoing simultaneous fMRI scanning. Researchers first marked stimuli appearing within each TR (repetition time), extracted word co-occurrence vectors corresponding to these stimuli as semantic representations for that TR, then built ridge regression prediction models using semantic representation vectors to predict activity in each brain voxel. If a voxel' s prediction correlation remained significant after multiple comparison correction, its activity was considered to contain semantic information, meaning it participated in semantic representation. Results showed that semantic information representation in the brain covered multiple brain regions including the medial prefrontal cortex, middle temporal gyrus, and temporoparietal junction, highly overlapping with the semantic network identified through meta-analysis [**?**, **?**]. These findings demonstrate that NLP-based semantic representation can be effectively applied to complex natural stimuli and further support the distributed representation view of semantics [**?**, **?**, **?**], where multiple brain regions jointly process and represent semantics rather than being concentrated in a single local area.

Furthermore, the quantification function of NLP techniques for lexical semantics enables researchers to examine semantic representation from finer-grained perspectives, broadening research scope. For example, Kivisaari et al. (2019) investigated the relationship between people' s representation of concepts and concept features. In the study, participants were presented with three feature words for each target concept (e.g., "a fruit," "peeled," "monkeys eat it"), and they needed to guess the corresponding concept (e.g., "banana") based on these features. Researchers decoded feature words or target words from brain voxel activity patterns and compared decoding accuracy rates for word vectors containing different information. Results showed that although participants only saw three feature words, adding all features of the target concept (including unpresented features) yielded the highest decoding accuracy, significantly higher than for presented feature words and the target concept alone, indicating that

the human brain constructs complete semantic representations of target objects using limited information fragments and activates other associated conceptual feature information.

3.2.2 The Influence of Contextual Information and Sentence-Level Semantic Representation

When investigating semantic representation in the brain, many studies present words or target stimuli in isolation, hoping to obtain semantic representations without interference from other information. However, semantic representation is dynamic [**?**, **?**], and the same word can express different meanings and produce different psychological feelings in different contexts. For example, the mental representation evoked by the word "women' s volleyball" differs from that of "Chinese women' s volleyball," with the latter potentially activating additional information such as pride and specific character images within the "Chinese" context. Research has shown that brain regions including the anterior temporal lobe and frontoparietal network integrate and update current semantic information [**?**, **?**, **?**, **?**, **?**], further demonstrating the dynamic nature of semantic representation. Context-independent experimental designs or static word vectors cannot fully capture semantic representations in rich contexts, especially when facing polysemy.

NLP techniques provide various deep language models capable of integrating context, such as ELMo [**?**, **?**], InferSent [**?**, **?**], and BERT, where semantic vectors for the same word can change with different contexts. Leveraging this characteristic, some researchers have compared representations of isolated words versus context-integrated words in the human brain [**?**, **?**]. In the experiment, each trial contained two sequentially presented English words, and participants needed to judge whether they were semantically related. Researchers first used the Word2Vec model to extract semantic vectors, which provide relatively fixed semantic representations for words unaffected by contextual words, thus reflecting isolated semantics. Simultaneously, for the same words, researchers also used the ELMo model to extract semantic vectors, which employs a bidirectional recurrent neural network structure and outputs word vectors that fully integrate contextual information (i.e., the preceding word). By using RSA to compare internal representational similarity between the human brain and language models for the same set of stimuli, researchers found that isolated semantic representation was primarily handled by the supramarginal gyrus, while context-dependent semantic representation was mainly associated with the left prefrontal cortex, angular gyrus, and ventral temporal lobe.

By using self-attention mechanisms to integrate contextual information, NLP techniques also provide metrics for representing sentence-level semantics (e.g., output vectors from the InferSent model or CLS vectors from BERT output). Sentence-level vector representations consider not only individual word semantics but also combinatorial relationships between words. In a recent study, participants viewed sentences composed of 4-9 words while undergoing fMRI

scanning. Researchers first used the InferSent model to extract sentence semantic representations, then established predictive relationships between sentence semantic features and brain activity patterns through ridge regression. Results showed that brain regions representing sentence meaning were distributed across extensive areas including the inferior frontal gyrus, middle frontal gyrus, superior temporal gyrus, middle temporal gyrus, and middle occipital gyrus [?, ?]. In another study, participants watched movies while undergoing fMRI scanning. Researchers segmented movies into multiple clips, provided text annotations for each clip (approximately 15 words per annotation), then used NLP models to convert annotations into semantic vectors as semantic features for movie clips, and finally predicted text annotation semantic features for each clip based on brain activity data. The study showed that brain activity patterns in the default mode network, language network, and occipital lobe could accurately predict clip semantic features and distinguish different clips [?, ?]. Consistent with these results, Acunzo et al. (2022) first trained a convolutional neural network for topic classification to enable model vectors to better capture topic information, then extracted output layer vectors from this model as sentence topic vector representations. Representational similarity analysis between topic vectors and brain activity revealed that the anterior temporal lobe, default mode network, and other regions participated in topic-level information representation, supporting the view that the default mode network has meaning construction functions involving abstraction and integration of long-term information [?, ?, ?].

3.2.3 Separating Syntax and Semantics

Successful communication of linguistic information depends not only on word semantics and background information provided by context but also on appropriate organizational structures between words, namely syntax. Classic syntactic research paradigms primarily adopt contrastive approaches attempting to isolate syntactic processing components, such as jabberwocky sentences replacing content words like nouns and adjectives with pseudowords [?, ?, ?], syntactic violations [?, ?, ?], syntactic adaptation [?, ?], and phrase combination [?, ?]. However, traditional syntactic processing research methods have limitations: for instance, brain regions for syntactic processing obtained from different tasks show considerable variation, and since semantics and syntax always co-occur, changing syntax without altering semantics is difficult [?, ?]. Therefore, syntactically scrambled sentences largely destroy semantic information, making it difficult for traditional experiments to isolate fine-grained syntactic processing [?, ?].

Word order structures in natural language texts contain rich linguistic information. Even without explicit syntactic relationships, NLP models with context integration capabilities learn syntactic relationships during training—for example, "I," "love," and "you" appear in the order "I love you" rather than "I you love." Deep language models (e.g., BERT) have approached or even surpassed human performance on various syntactic tasks including subject-verb agreement and re-

flexive pronoun anaphora [?, ?, ?], indicating their ability to accurately acquire syntactic information from text. Adopting the "subtraction" approach from experimental design, NLP models can be used to separately extract syntactic and semantic information from sentences, stripping syntactic information from vectors to investigate brain regions processing syntactic information [?, ?, ?, ?]. Research results show that both bilateral temporal lobes and inferior frontal gyri process syntactic information, with brain region distributions similar to previous experimental studies [?, ?].

Recently, some researchers have used feature elimination methods to more finely separate syntactic information (e.g., part of speech, named entities, word dependencies, semantic roles) and investigate various syntactic processes when participants listen to stories [?, ?]. Results showed that although brain region distributions for different syntactic features had minor differences, the distributed regions were roughly the same, concentrated in semantic network areas such as the superior temporal gyrus, middle temporal gyrus, and angular gyrus [?, ?].

NLP models can effectively separate semantic and syntactic information and enable investigation of brain processing mechanisms in less constrained natural tasks, indicating great potential for brain representation research [?, ?, ?]. However, current research using NLP models to investigate brain syntactic processing is limited, and the syntactic processing brain regions identified are broader than those found with traditional research methods. Whether this phenomenon reflects the true distributed processing mechanism of syntactic information or stems from errors in mapping between NLP models and brain imaging data requires further analysis in future research.

3.2.4 Representation of Discourse Topic Information and Discourse Semantic Structure

Discourse (paragraph) comprehension builds upon semantic analysis of words and sentences, forming representations of core topic information (or situation models) by identifying semantic structural relationships between different parts of discourse and integrating contextual information [?, ?]. Traditional experimental methods typically compare intact discourse with scrambled materials [?, ?, ?, ?], but scrambled materials increase memory and integration difficulty for participants, so detected differences may not be entirely driven by processing specific to discourse-level semantics. Moreover, this method does not quantify discourse information, making it difficult to measure semantic distances and relationships between discourses, and is unsuitable for research using different discourse materials.

In recent years, researchers have begun using NLP techniques to model discourse semantics and investigate human brain processing and representation of continuous natural language stimuli (e.g., stories or movies). A recent study combined fMRI technology with LSA methods to explore how complex narrative information presented in different modalities is represented in the human

brain [?, ?]. During fMRI scanning, one group of participants watched silent films while another group listened to voice narrations corresponding to the film content. After scanning, participants described story content in their own words. Researchers used LSA for semantic analysis and found that regardless of whether participants watched silent films or listened to voice narrations, higher semantic similarity in their described content was associated with higher similarity in neural activity in the default mode network and executive control network. This result reveals the function of the default mode network (DMN) in cross-modal representation of thematic semantic information. Another study examined the consistency of brain representation of topic information between speech production and comprehension processes [?, ?]. During fMRI scanning, participants orally described a series of topics and listened to other topics described by another participant. Researchers used LSA to calculate semantic distances between descriptions and computed brain representational dissimilarity matrices for both speech comprehension and production tasks, then calculated similarity between semantic and brain representational dissimilarity matrices (RSA analysis). Results showed that activity patterns in bilateral extensive brain regions including the inferior frontal gyrus, medial prefrontal cortex, temporal pole, middle temporal gyrus, angular gyrus, and precuneus were associated with semantic content in both speech comprehension and production. This study, the first to analyze discourse-level semantics in speech production, revealed a shared network for high-level discourse semantic information representation between speech production and comprehension processes. These studies, through analysis of discourse-level semantic information, further support the role of the default mode network in meaning construction [?, ?, ?].

Discourse materials can also be investigated from network topology properties to explore how semantic structure affects brain processing, learning, and memory. In natural stimuli such as text and video, sentences and events are interconnected within a theme—for example, a story typically unfolds around several core topic sentences or plots. Using semantic similarity as edge weights to construct topological networks for discourse can reflect discourse semantic organizational structure information. Some researchers have investigated the relationship between movie narrative rhythm and audience ratings [?, ?], using semantic similarity between adjacent clips as an indicator of plot development rate, with higher similarity between adjacent clips indicating slower plot development. Results showed that movies with slow beginnings and slightly faster endings received higher ratings, indicating that discourse semantic structure affects people's feelings and engagement. Another recent brain imaging study examined how discourse semantic structure affects memory [?, ?]. Researchers segmented video clips, extracted semantic vectors from text descriptions of each clip using NLP techniques, and constructed a video semantic structure topological network with clips as nodes and semantic similarities between clips as edge weights. Results showed that clips with higher centrality (reflecting stronger associations with other nodes) produced better memory effects and elicited stronger activation and higher inter-subject consistency in brain regions related to episodic

memory (the default mode network), indicating that human brain processing and memory of events are related to their positions in semantic organizational structure.

These results demonstrate that discourse semantic organizational structure affects people' s subjective feelings, memory effects, and brain activity. However, current research using NLP on brain semantic representation mostly approaches from stimulus encoding, paying less attention to semantic organizational structure and relationships in continuous stimuli. Future research could start from semantic structure in natural stimuli to further explore its associations with brain processing, learning, and memory effects, such as the neural basis of conspiracy theory and rumor identification [**?**, **?**] and narrative preferences [**?**, **?**].

3.2.5 Summary

The use of NLP technology transforms language from symbolic representation to vector representation, overcoming to some extent the difficulties of discreteness, quantification, and unified representation of words, making semantic computation and comparison possible. Meanwhile, multivariate analysis methods such as representational similarity analysis and linear regression bridge different modalities of data. With the development of deep language models, NLP models can now integrate contextual information into vector representations, improving language representation precision and enabling real-time characterization of semantic dynamics across different contextual backgrounds. Based on this, researchers use word vectors extracted by NLP as semantic representations, reducing the need for artificial control of stimulus materials or experimental tasks, and making the exploration of semantic brain representation no longer dependent on comparisons between different types of stimuli or processing tasks. Additionally, NLP, as a computational language model, has high flexibility—inputting different types of text yields corresponding information. Researchers can analyze information types or processing characteristics of certain brain regions by comparing how well model vector representations of different text types (e.g., context-containing versus context-free word vectors) match brain representations [**?**, **?**], such as the human brain' s mechanism for predicting future words [**?**, **?**, **?**] and the influence of prior beliefs on text comprehension [**?**, **?**]. By shifting the focus of experimental design from brain activity to computational models, NLP technology can be used to separate different information components and effectively reduce requirements for participant numbers and experiments. Finally, the use of natural stimuli and low-constraint tasks is gradually becoming a trend in brain imaging research [**?**, **?**, **?**], yet traditional psychological experimental methods struggle to track continuously input word semantics and integrate previous contextual information into current words. NLP technology provides modeling methods for representing multi-level semantic information including characters, words, sentences, and discourse, playing an increasingly important role in exploring the neural basis of natural language processing.

Using NLP technology to extract semantic features of stimuli and establish

mapping relationships with brain activity, recent researchers have consistently observed that neural structures related to semantic representation are widely distributed across multiple brain regions including the frontal, temporal, and occipital lobes. This result is not entirely consistent with conclusions from traditional psychological experimental methods and brain lesion patients that revealed localized brain regions for semantic representation. One possible reason is that language models trained on large text corpora have more fully captured multiple semantic information of linguistic symbols, while traditional psychological experiments using specific tasks (e.g., semantic association judgment) selectively activate one aspect of linguistic symbol semantics, thus previously detecting only partial brain region involvement. Notably, many theoretical models have also proposed that neural representations of semantic memory are distributed across extensive brain regions including sensory-motor and association cortices [**?**, **?**, **?**, **?**, **?**]. For example, the "hub-and-spoke" theory of conceptual representation [**?**, **?**, **?**] proposes that cross-modal linguistic and non-linguistic experiences constitute the core components of concepts (the hub), primarily represented and integrated by the anterior temporal lobe, while the initial source information appearing during concept acquisition (the spokes, including visual, auditory, emotional valence, etc.) is distributed across various modality-specific cortices. Additionally, dual coding theory divides knowledge representation into two categories: sensorimotor-derived systems and language-derived systems, where the knowledge representation system supporting sensorimotor coding is mainly distributed across modality-specific sensory-motor cortices and association cortices, while the system supporting linguistic coding is mainly distributed in the dorsal anterior temporal lobe (dATL) and its extended regions (including classic language brain regions such as the inferior frontal gyrus and middle temporal gyrus). The broad semantic-sensitive brain regions revealed by NLP technology suggest that vector spaces representing semantics may simultaneously capture both abstract, cross-modal components and modality-specific components of natural language, yet establishing definitive associations between these findings and cognitive theoretical models still faces numerous challenges (for more in-depth discussion of this issue, please refer to: [**?**, **?**, **?**]).

4. Summary and Outlook

Compared with traditional psychological experimental methods, using natural language processing (NLP) technology to characterize semantics has several advantages: (1) It enables objective quantification and computation of semantic information at multiple levels including words, sentences, and discourse, providing metrics for semantics; (2) It can integrate contextual information and adjust word vector outputs according to context, thereby achieving more accurate representation of semantics in context; (3) Word vectors output by NLP models contain rich information, and through ablation experiments or inputting different types of stimuli, researchers can extract or remove certain information (e.g., syntactic information) to examine brain semantic representation from different information perspectives; (4) Word vector acquisition is fast and convenient,

less subject to subjective factors, and can greatly reduce material rating costs.

Through methods such as representational similarity and linear regression, researchers have attempted to use semantic information extracted from language models to explain changes in brain activity, achieving many new discoveries regarding distributed semantic representation, the influence of contextual information on semantic representation, separation of syntactic and semantic processing regions, and discourse semantic representation.

However, natural language processing technology also has certain limitations when answering questions about language cognition and its neural mechanisms. First is the interpretability issue of NLP models. In recent years, language models based on neural networks and deep learning have become increasingly complex and large in internal structure—for example, the recent GPT-3 model has 175 billion parameters [**?**, **?**]. Despite good performance on language tasks, the huge parameter count and complex structure make model interpretability poor: What linguistic features are reflected in model output word vectors? What key steps does the model use to obtain these features? These questions currently have no definitive answers. While model comparison approaches (e.g., eliminating or preserving contextual information, using random vectors instead of word vectors) can be used to investigate brain processing of certain information, low interpretability still somewhat limits the explanatory power and application potential of NLP in language cognition research. Second, the number and types of models are growing rapidly, with differences across models in training materials, network architecture, parameter count, and training tasks, leading to different output word vectors. When establishing mapping relationships between word vectors and brain activity, the sources of performance differences between models become unclear. Even using the same pre-trained model to obtain identical model parameters, issues such as model sampling error remain. Additionally, the construction of NLP models differs from how humans acquire semantics, and their internal computational and processing mechanisms may fundamentally differ from the human brain. Human language acquisition is a process of continuous multimodal interaction with the world environment, whereas most current mainstream NLP models have only text modality and cannot learn new knowledge or change existing concepts based on just a few feedback instances like humans. On the other hand, despite increasingly large training corpora and complex structures, NLP models still perform poorly on advanced language tasks such as logical reasoning and knowledge transfer, raising the question of whether NLP truly learns language. Therefore, the extent to which NLP models can explain semantic representation mechanisms in the human brain requires deeper future research. Given these limitations, when applying language models to extract stimulus features, researchers need to select appropriate models according to research questions, test model effectiveness through experimental design, and interpret results cautiously.

Notably, NLP models are not always the only or optimal solution for semantic representation. Other current psychological semantic representation methods

have also achieved good performance in some cases and have strong interpretability. For example, feature listing methods can intuitively reflect the salience of different concept features in memory [**?**, **?**]; feature rating methods can obtain attribute strengths of concepts across multiple dimensions (e.g., perception, emotion) and perform similarity calculations using distributed representations [**?**, **?**]; network models can clearly reflect hierarchical and relational structures between concepts [**?**, **?**, **?**]. NLP models trained purely on text may not necessarily fully capture human semantic knowledge and processing characteristics (e.g., reasoning, association, multimodality). For example, recent research on conceptual semantic brain representation found that compared with NLP models, feature ratings based on experiential attributes showed higher representational similarity with the brain, and after controlling for shared information using partial correlation, experiential attributes still showed significant representational similarity with the brain, while NLP models did not, indicating that human brain representation of concepts contains multimodal information not yet learned by NLP models [**?**, **?**, **?**]. Therefore, there is no absolute superiority between NLP models and traditional psychological semantic representation methods; they provide complementary information and functions [**?**, **?**]: In small-scale corpora, although traditional methods are coarse-grained, their high interpretability helps verify research theories and hypotheses; in large-scale corpora and natural stimuli, although the low interpretability of NLP models makes vector dimension meanings unclear, they can conveniently obtain contextualized semantic representations and examine different information content through model comparisons.

Next, researchers can further expand the application of NLP technology in neurolinguistics from the following aspects:

(1) **Introducing graph-based semantic representation methods.** In addition to text representation methods based on distributed hypothesis, graph models are also mature techniques in NLP for representing text relationships (e.g., knowledge graphs). In graph models, network nodes represent linguistic elements (words, concepts, entities, sentences, discourse, etc.), and network edges represent relationships between linguistic elements. Taking knowledge graphs as an example, graph model construction fully utilizes attribute relationships between linguistic elements, linguistic prior knowledge, and world knowledge, offering higher interpretability than neural network models with clear semantic relationships that facilitate commonsense reasoning tasks. However, the data structures used by graph models to represent semantics are relatively complex, making it difficult to directly model brain activity data using graph model semantic representations. Researchers can adopt indirect approaches, extracting semantic relationship or distance information from graph models, then using RSA and other methods to examine brain processing of semantic relationships. Using WordNet as an example, this database organizes words in tree structures based on semantic relationships (e.g., hierarchical relationships). Semantic distance between two words in WordNet can be

measured by the shortest path connecting them [**?**, **?**, **?**, **?**]. For example, reaching the node for "mouse" from the node for "cat" requires the following path: cat–feline–carnivore–placental mammal–rodent–mouse, so the relational distance between these two words is 5.

(2) **Applying multimodal fusion deep language models.** In natural communication contexts, people' s information processing and understanding often integrate multiple modalities such as sound, images, and text, and processing single concepts often extracts multimodal information [**?**, **?**]. However, traditional experimental methods and pure text-based NLP models struggle to fuse and quantify multimodal information and cannot comprehensively describe content of human brain concept representation [**?**, **?**, **?**]. The artificial intelligence field has developed multimodal fusion deep semantic representation methods [**?**, **?**, **?**, **?**]. Using multimodal language models can further investigate brain processing mechanisms for different modality information, such as the representation distribution and patterns, role status, and integration methods and degrees of language-based versus experience-based information [**?**, **?**, **?**] in the brain.

(3) **Using language models to assess language abilities in special populations.** For example, performing text analysis on language production from normal individuals and patients with aphasia (or autism, schizophrenia, etc.) to obtain features such as semantic categories, semantic ambiguity, word frequency distribution, and semantic structure [**?**, **?**, **?**]. Building classification or prediction models based on these features can help improve the accuracy or acceptability of language ability and disease assessment [**?**, **?**, **?**] and reduce time and labor costs required for evaluation.

(4) **Using brain activity data to enhance understanding or improve deep language models.** Current deep language models can complete various language tasks, yet people still lack clear understanding of their internal implementation mechanisms. The human brain is the only processing system in the world that can truly understand natural language. One approach to understanding deep models is to compare them with the human brain. Some studies have begun using the "brain-likeness" of deep models to infer internal operating mechanisms or explain differences between models. For example, in one study, researchers investigated the contextual integration capabilities of different language models and different hidden layers within the same model [**?**, **?**]. Researchers used fMRI to collect brain activity while participants read stories (with each word presented separately on screen), simultaneously extracting vector representations for each word in the story from each hidden layer of different NLP models, then calculating the prediction degree of model output word vectors on activity in multiple important language brain areas through ridge regression and classification tasks. Results showed that when shorter contexts (fewer than 10 words) were used to compute word vectors, middle

layers of BERT and Transformer T-XL models predicted brain activity better than shallower input layers, reflecting hidden layers' contextual integration capabilities. When contextual information exceeded 10 words, BERT' s brain activity prediction effect declined with increasing context word count, while Transformer T-XL' s prediction effect continued to show a slow upward trend. Researchers speculated that the context length corresponding to optimal brain activity prediction might reflect the model' s (or hidden layer' s) ability to integrate contextual information, showing that Transformer T-XL is better at integrating long-distance contextual information than BERT, which was indeed one of Transformer T-XL' s original design goals. Similar work has also found significant positive correlations between NLP models' language task performance and brain activity prediction ability [**?**, **?**, **?**]. Furthermore, some researchers have fine-tuned models and found that improving brain activity prediction ability (making models more "brain-like") significantly enhanced model performance on multiple language tasks [**?**, **?**, **?**].

These studies indicate that comparing deep language models with human brain cognitive and neural processing has great potential for understanding and even improving deep language models. However, due to the covert nature of human thinking and limitations of current brain imaging technology in temporal and spatial resolution as well as low signal-to-noise ratio, "brain-like" analysis or investigation of internal cognitive mechanisms of NLP models still requires using strict experimental controls and prior knowledge to constrain results or cooperating with other model interpretation methods to make inferences [**?**, **?**].

References

王少楠, 丁鼎, 林楠, 张家俊, 宗成庆. (2022a). 语言认知与语言计算——人与机器的语言理解. 中国科学: 信息科学, 52(10), 1748-1774. https://doi.org/10.1360/SSI-2021-0100

王少楠, 张家俊, 宗成庆. (2022b). 基于语言计算方法的语言认知实验综述. 中文信息学报, 36(4),

赵京胜, 宋梦雪, 高祥, 朱巧明. (2022). 自然语言处理中的文本表示研究. 软件学报, 33(1), 102-128. https://doi.org/10.13328/j.cnki.jos.006304

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, 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. https://doi.org/10.1523/JNEUROSCI.1152-20.2021

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. S. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict

brain activity using stimulus-model-similarities. *NeuroImage*, 128, 44–53. https://doi.org/10.1016/j.neuroimage.2015.12.035

Batterink, L., & Neville, H. J. (2013). The human brain processes syntax in the absence of conscious awareness. *Journal of Neuroscience*, 33(19), 8528–8533. https://doi.org/10.1523/JNEUROSCI.0618-13.2013

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(6), https://doi.org/10.1162/153244303322533223

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, 25(10), 883–895. https://doi.org/10.1016/j.tics.2021.07.006

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

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

Bonnici, H. M., Richter, F. R., Yazar, Y., & Simons, J. S. (2016). Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. *Journal of Neuroscience*, 36(20), 5462–5471. https://doi.org/10.1523/JNEUROSCI.4310-15.2016

Branzi, F. M., Humphreys, G. F., Hoffman, P., & Lambon Ralph, M. A. (2020). Revealing the neural networks that extract conceptual gestalts from continuously evolving or changing semantic contexts. *NeuroImage*, 220, 116802. https://doi.org/10.1016/j.neuroimage.2020.116802

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467–480. https://aclanthology.org/J92-4003

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://dl.acm.org/doi/10.5555/3495724.3495883

Bruffaerts, R., De Deyne, S., Meersmans, K., Liuzzi, A. G., Storms, G., & Vandenberghe, R. (2019). Redefining the resolution of semantic knowledge in the brain: Advances made by the introduction of models of semantics in neuroimaging. *Neuroscience & Biobehavioral Reviews*, 103, 3–13. https://doi.org/10.1016/j.neubiorev.2019.05.015

Carota, F., Nili, H., Pulvermuller, F., & Kriegeskorte, N. (2021). Distinct fronto-temporal substrates of distributional and taxonomic similarity among

words: Evidence from RSA of BOLD signals. *NeuroImage*, 224, 117408. https://doi.org/10.1016/j.neuroimage.2020.117408

Caucheteux, C., Gramfort, A., & King, J. R. (2021a). Disentangling syntax and semantics in the brain with deep networks. *Proceedings of the 38th International Conference on Machine Learning*, 139, 1336–1348. https://proceedings.mlr.press/v139/caucheteux21a.html

Caucheteux, C., Gramfort, A., & King, J. R. (2021b). Long-range and hierarchical language predictions in brains and algorithms. *arXiv*. https://doi.org/10.48550/arXiv.2111.14232

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton. https://doi.org/10.1515/9783112316009

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, 23(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670–680. https://doi.org/10.18653/v1/D17-1070

Cooper, K. E., & Nisbet, E. C. (2016). Green narratives: How affective responses to media messages influence risk perceptions and policy preferences about environmental hazards. *Science Communication*, 38(5), 626–654. https://doi.org/10.1177/1075547016666843

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2), 163–201. https://doi.org/10.1037/0096-3445.132.2.163

Day, M., Dey, R. K., Baucum, M., Paek, E. J., Park, H., & Khojandi, A. (2021). Predicting severity in people with aphasia: A natural language processing and machine learning approach. *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, 2299–2302. https://doi.org/10.1109/embc46164.2021.9630694

de Boer, J. N., Voppel, A. E., Begemann, M. J. H., Schnack, H. G., Wijnen, F., & Sommer, I. E. C. (2018). Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, https://doi.org/10.1016/j.neubiorev.2018.06.008

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Dubova, M. (2022). Building human-like communicative intelligence: A grounded perspective. *Cognitive Systems Research*, 72, 63–79. https://doi.org/10.1016/j.cogsys.2021.12.002

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38, 189–230. https://doi.org/10.1002/aris.1440380105

Dupre la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, https://doi.org/10.1016/j.neuroimage.2022.119728

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fedorenko, E., Nieto-Castanon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, 50(4), 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., . . . Seidenberg, M. S. (2016a). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, 26(5), 2018–2034. https://doi.org/10.1093/cercor/bhv020

Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S., & Binder, J. R. (2016b). Heteromodal cortical areas encode sensory-motor features of word meaning. *Journal of Neuroscience*, 36(38), 9763–9769. https://doi.org/10.1523/JNEUROSCI.4095-15.2016

Fernandino, L., Tong, J. Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences of the United States of America*, 119(6). https://doi.org/10.1073/pnas.2108091119

Finn, E. S., & Bandettini, P. A. (2021). Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage*, https://doi.org/10.1016/j.neuroimage.2021.117963

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer' s disease in narrative speech. *Journal of Alzheimer' s Disease*, 49(2), 407–422. https://doi.org/10.3233/jad-150520

Gao, Z., Zheng, L., Gouws, A., Krieger-Redwood, K., Wang, X., Varga, D., ···& Jefferies, E. (2023). Context free and context-dependent conceptual representation in the brain. *Cerebral Cortex*, 33(1), 152–166. https://doi.org/10.1093/cercor/bhac058

Goldberg, Y. (2019). Assessing BERT' s syntactic abilities. *arXiv*. https://doi.org/10.48550/arXiv.1901.05287

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., . . . Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Gonzalez, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuan, A., Belloch, V., & Avila, C. (2006). Reading cinnamon activates olfactory brain regions. *NeuroImage*, 32(2), 906–912. https://doi.org/10.1016/j.neuroimage.2006.03.037

Graves, A., Mohamed, A. r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, 37(1), 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5), https://doi.org/10.1080/23273798.2018.1499946

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

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Heusser, A. C., Fitzpatrick, P. C., & Manning, J. R. (2021). Geometric models reveal behavioural and neural signatures of transforming experiences into memories. *Nature Human Behaviour*, 5(7), 905–919. https://doi.org/10.1038/s41562-021-01051-6

Hobbs, J. R. (1977). Pronoun resolution. *ACM SIGART Bulletin*(61), https://doi.org/10.1145/1045283.1045292

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Humphreys, G. F., Lambon Ralph, M. A., & Simons, J. S. (2021). A unifying account of angular gyrus contributions to episodic and semantic cognition. *Trends in Neurosciences*, 44(6), 452–463. https://doi.org/10.1016/j.tins.2021.01.006

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

Jain, S., & Huth, A. G. (2018). Incorporating context into language encoding models for fMRI. *Advances in Neural Information Processing Systems*, 31, 6629–6638.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, 48(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Kivisaari, S. L., van Vliet, M., Hulten, A., Lindh-Knuutila, T., Faisal, A., & Salmelin, R. (2019). Reconstructing meaning from bits of information. *Nature Communications*, 10(1), 927. https://doi.org/10.1038/s41467-019-08848-0

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55, 167–179. https://doi.org/10.1016/j.conb.2019.04.002

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, https://doi.org/10.3389/neuro.06.004.2008

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9), 1449–1477. https://doi.org/10.1109/jproc.2015.2460697

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Laurino Dos Santos, H., & Berger, J. (2022). The speed of stories: Semantic progression and narrative success. *Journal of Experimental Psychology: General*, 151(8), https://doi.org/10.1037/xge0001171

Law, R., & Pylkkanen, L. (2021). Lists with and without syntax: A new approach to measuring the neural processing of syntax. *Journal of Neuroscience*, 41(10), https://doi.org/10.1523/JNEUROSCI.1179-20.2021

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. https://doi.org/10.1038/44565

Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, 13(1), 4235. https://doi.org/10.1038/s41467-022-31965-2

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lu, H., Zhou, Q., Fei, N., Lu, Z., Ding, M., Wen, J., . . . Wen, J.-R. (2022). Multimodal foundation models are better simulators of the human brain. *arXiv*. https://doi.org/10.48550/arXiv.2208.08263

Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2020). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature Human Behaviour*, 5(3), 327–336. https://doi.org/10.1038/s41562-020-00985-7

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., . . . Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America*, 113(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113

Matchin, W., Brodbeck, C., Hammerly, C., & Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping*, 40(2), 663–678. https://doi.org/10.1002/hbm.24403

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schutze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences of the United States of America*, 117(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

Mcculloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. https://doi.org/10.1007/bf02478259

Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, 8(43), eabq3668. https://doi.org/10.1126/sciadv.abq3668

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J. H., & Khudanpur, S. (2010). Recurrent neural network based language model. *11th Annual Conference of the International Speech Communication Association*, 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *arXiv*. https://doi.org/10.48550/arXiv.1310.4546

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880), 1191–1195. https://doi.org/10.1126/science.1152876

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti Di Oleggio Castello, M., . . . Haxby, J. V. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, 27(8), 4277–4291. https://doi.org/10.1093/cercor/bhx138

Nevler, N., Ash, S., McMillan, C., Elman, L., McCluskey, L., Irwin, D. J., . . . Grossman, M. (2020). Automated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders. *Neurology*, 95(12), E1629–E1639. https://doi.org/10.1212/wnl.0000000000010366

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage*, https://doi.org/10.1016/j.neuroimage.2018.09.010

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3), 255–287. https://doi.org/10.1037/h0084295

Patel, T., Morales, M., Pickering, M. J., & Hoffman, P. (2022). A common neural code for meaning in discourse production and comprehension. *bioRxiv*. https://doi.org/10.1101/2022.10.15.512349

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12), 976–987. https://doi.org/10.1038/nrn2277

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Petersson, K.-M., Folia, V., & Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language*, 120(2), https://doi.org/10.1016/j.bandl.2010.08.003

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage*, 10(1), 15–35. https://doi.org/10.1006/nimg.1999.0441

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates

using GLMsingle. *eLife*, 11, e77599. https://doi.org/10.7554/elife.77599

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9), https://doi.org/10.1016/j.tics.2013.06.004

Pulvermüller, F., Harle, M., & Hummel, F. (2001). Walking or talking? Behavioral and neurophysiological correlates of action verb processing. *Brain and Language*, 78(2), 143–168. https://doi.org/10.1006/brln.2000.2390

Pulvermüller, F., Kherif, F., Hauk, O., Mohr, B., & Nimmo-Smith, I. (2009). Distributed cell assemblies for general lexical and category-specific semantic processing as revealed by fMRI cluster analysis. *Human Brain Mapping*, 30(12), 3837–3850. https://doi.org/10.1002/hbm.20811

Pulvermüller, F., Lutzenberger, W., & Preissl, H. (1999). Nouns and verbs in the intact brain: Evidence from event-related potentials and high-frequency cortical responses. *Cerebral Cortex*, 9(5), 497–506. https://doi.org/10.1093/cercor/9.5.497

Pylkkanen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, 366(6461), 62–66. https://doi.org/10.1126/science.aax0050

Quoc, L., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, 32, 1188–1196.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. https://doi.org/10.1145/361219.361220

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

Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, 32, 14123–14133.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension—an fMRI study. *Cerebral Cortex*, 22(7), 1662–1670. https://doi.org/10.1093/cercor/bhr249

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1), 12141. https://doi.org/10.1038/ncomms12141

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., & Margulies, D. S. (2021). The default mode network in cognition: A topographical perspective. *Nature Reviews Neuroscience*, 22(8), 503–513. https://doi.org/10.1038/s41583-021-00474-4

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. *Behavior Research Methods*, 51(4), 1717–1736. https://doi.org/10.3758/s13428-019-01217-8

Sun, X., Yang, D., Li, X., Zhang, T., Meng, Y., Qiu, H., . . . Li, J. (2021). Interpreting deep learning models in natural language processing: A review. *arXiv*. https://doi.org/10.48550/arXiv.2110.10470

Sundermeyer, M., Schluter, R., & Ney, H. (2012). LSTM neural networks for language modeling. *13th Annual Conference of the International Speech Communication Association*, 194–197.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27.

Tikochinski, R., Goldstein, A., Yeshurun, Y., Hasson, U., & Reichart, R. (2021). Fine-tuning of deep language models as a computational framework of modeling listeners' perspective during language comprehension. *bioRxiv*. https://doi.org/10.1101/2021.11.22.469596

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 14954–14964.

Tong, J., Binder, J. R., Humphries, C., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A distributed network for multimodal experiential representation of concepts. *Journal of Neuroscience*, 42(37), 7121–7130. https://doi.org/10.1523/JNEUROSCI.1243-21.2022

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Vodrahalli, K., Chen, P. H., Liang, Y., Baldassano, C., Chen, J., Yong, E., . . . Arora, S. (2018). Mapping between fMRI responses to

movies and their natural language annotations. *NeuroImage*, 180, 223–231. https://doi.org/10.1016/j.neuroimage.2017.06.042

Wang, S., Zhang, J., Lin, N., & Zong, C. (2020). Probing brain activation patterns by dissociating semantics and syntax in sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 9201–9208. https://doi.org/10.1609/aaai.v34i05.6457

Wang, S., Zhang, J., & Zong, C. (2018). Associative multichannel autoencoder for multimodal word representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 115–124. https://doi.org/10.18653/v1/D18-1011

Wang, S., Zhang, J., Lin, N., & Zong, C. (2018). Investigating inner properties of multimodal representation and semantic compositionality with brain-based componential semantics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30, 5964–5972. https://doi.org/10.1609/aaai.v32i1.12032

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., . . . Bi, Y. (2018). Organizational principles of abstract words in the human brain. *Cerebral Cortex*, 28(12), https://doi.org/10.1093/cercor/bhx283

Warburton, E., Wise, R. J., Price, C. J., Weiller, C., Hadar, U., Ramsay, S., & Frackowiak, R. S. (1996). Noun and verb retrieval by normal subjects studies with PET. *Brain*, 119, 159–179. https://doi.org/10.1093/brain/119.1.159

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS One*, 9(11), e112575. https://doi.org/10.1371/journal.pone.0112575

Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 833–844. https://doi.org/10.18653/v1/D19-1077

Wurm, M. F., & Caramazza, A. (2019). Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nature Communications*, 10(1), 289. https://doi.org/10.1038/s41467-018-08084-y

Xu, C., Zhang, Y., Zhu, G., Rui, Y., Lu, H., & Huang, Q. (2008). Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 10(7), 1342–1355. https://doi.org/10.1109/Tmm.2008.2004912

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, 23(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: Where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, 22(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv*. https://doi.org/10.48550/arXiv.1702.01923

Zhang, X., Wang, S., Lin, N., Zhang, J., & Zong, C. (2022). Probing word syntactic representations in the brain by a feature elimination method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 11721–11729. https://doi.org/10.1609/aaai.v36i10.21427

Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv*, https://doi.org/10.48550/arXiv.1510.03820

Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., . . . Yuan, N. J. (2022). Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 1–20. https://doi.org/10.1109/tkde.2022.3224228

Note: Figure translations are in progress. See original paper for figures.