Beyond fluency: understanding in large language models and analogies with the human brain.

Diego Calanzone

Department of Information Engineering and Computer Science, University of Trento

Abstract

A recent paradigm shift in natural language processing allowed to successfully apply language models in multiple fields e.g. in robotics, computational biology and automated vision. The possibility for these algorithms to achieve or surpass human performance further pushed the question: to what extent do they process information similarly to the human brain? Studies in neuroscience and AI have been conducted in an isolated way, only in the recent years brain activity patterns have been shown to correlate with computations in neural networks. This article firstly aims at reviewing the most crucial findings, suggesting that studying the behavior of neural networks helps understanding more the brain. Further discussion and experiments are dedicated to testing the knowledge of language models: what features drives their understanding, how they perform predictions and to what extent they can reason and learn to execute algorithms. Findings suggest that neural networks can learn abstract concepts applicable to problems, but their logical abilities are limited either by the training procedure or the design of the network; such limitations may underlie divergences with the human brain.

Beyond fluency: understanding in large language models and analogies with the human brain.

Contents

Computational representations of language

Language is a human construct used to convey meaning. In computational linguistics theory, Chomsky (1956) defines language as a set of units of meaning, e.g. the words in a vocabulary, and rules to compose them into sequences. Understanding and modeling the criteria to construct language sentences has been a major question in research– a puzzling finding is that the distribution of human spoken language approximately follows a simple mathematical relation, namely the Zipf's Power Law. In a Zipfian distribution, given words ranked in descending order by frequency, the first term occurs n-times more frequently than the n-th most common term. As a result, there is a concentration of few highly occurring words against a large set of sparse terms: functional words such as articles or prepositions fall in the first category, e.g. "the" is the most frequent, accounting for 7% of word occurrencies in the Brown Corpus of American English text; content words such as "democracy" belong to the latter, as reported by Stephen and Ramazan (2010). The reason why a complex process such as speech follows such simple mathematical rule has been a subject of research for the past 70 years.

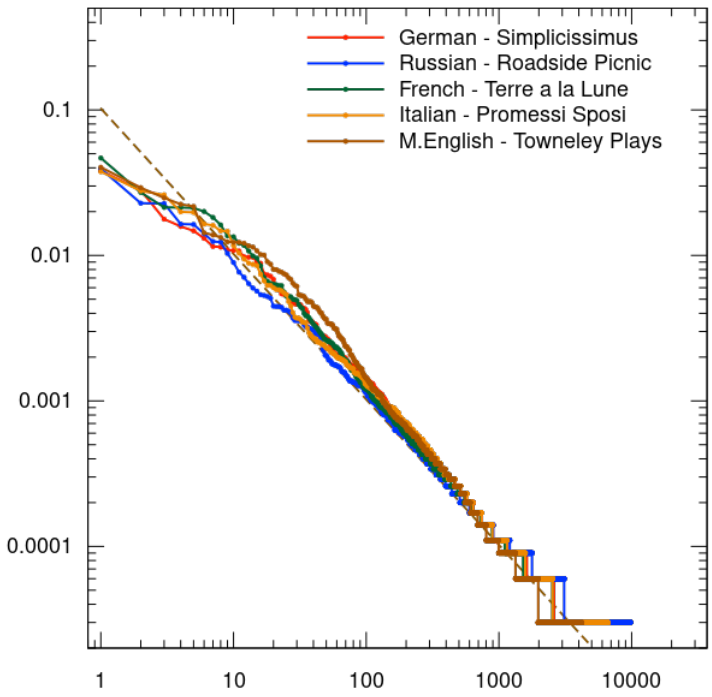

Figure 1. Zipf law on several languages, Wikipedia (2023).

Distributional semantics and sparse repesentations

The use of word frequency to extract features from language is at the base of computational models for linguistics. Arguably, the simplest technique to represent a body of text is to count the words occurring in it: a **bag of words (BoW)** consists in an array of word counts for a specific document, where the cardinality of this sequence (the "embedding") corresponds to the size of a universal vocabulary.

```
sentence 1: "Paris is in France"
sentence 2: "Rome is in Italy"
BoW 1: {"Rome": 0, "Paris": 1, "is": 1, "in": 1, "France": 1, "Italy": 0}
BoW 2: {"Rome": 1, "Paris": 0, "is": 1, "in": 1, "France": 0, "Italy": 1}
```

Example 1. Bag of words encoding for two example sentences.

Similarly, words can be represented by a vector of co-occurrence counts with respect to all the other terms in the vocabulary: this is defined as the term-term matrix or **co-occurrence matrix**. While both techniques have been applied in research with success, major problems lead to a paradigm shift: document/word representations are highly dimensional and vocabulary-specific, such vectors have been found to be highly sparse and thus inefficient for computation.

	Rome	Paris	is	in	France	Italy
Rome	1	0	1	0	0	0
Paris	0	1	1	0	0	0
is	0	0	1	1	0	0
in	0	0	0	1	1	1
France	0	0	0	0	1	0
Italy	0	0	0	0	0	1

Table 1. Co-occurrence matrix representation of the sentences from Example 1.

The above mentioned approach can be generalized in probability theory by

considering a language sentence as a stochastic process with words as random variables, the probability for a sequence to occur is modeled by the chain rule:

$$p(w_1, ..., w_n) = \prod_{i}^{n} p(w_i | w_{1,...,i-1}) \tag{1}$$

That is the probability of a word to follow in a sentence depends by the preceeding ones. This concept can be simplified with the Markov assumption: the conditioning sequence for a word is only a portion (the latest) of the preceeding terms. Words co-occurrence can be thus seen as a special case in which the probability of a word w_i is governed by the preceeding one w_j, a cell entry c_{ij} corresponds to the count for the row word to preceed the column word; the probability for w_i to follow w_j, a **bigram**, is defined as the fraction of occurrences of the two words over the occurrences of the preceding word:

$$p(w_i | w_j) = \frac{c_{ij}}{c_j} \tag{2}$$

This approach to language is defined as **language modeling**: human language is assumed to follow a distribution that governs the sequences of function and content words. More recent research in statistical learning and artificial intelligence propose the use of neural networks as non-parametric models for language modeling. At the center of this paradigm shift is the idea to move from discrete to continuous and compressed word representations: terms are mapped to vectors in a highly-dimensional geometrical space, which are subject to algebraic properties.

Neural networks and continuous embeddings

Mikolov, Chen, Corrado, and Dean (2013) implement the idea of words as geometrical points in what has been considered a turning point in natural language processing: a neural network is employed to predict the probability of co-occurrence between two words within a context window. Word embeddings consist in the vectors of logits extracted from the penultimate layer of the network, before predicting the single probability value; vectors with fixed cardinality and continuous values represent words

in a geometrical space. The authors show interesting properties in linearly composing these representations, e.g. with a mathematical analogy they compose high level concepts to infer new words.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter

Table 2. Examples extracted from inference with word2vec from Mikolov et al. (2013).

It is fundamental to mention that with this approach the representation for a word is unique and it does not depend on the surrounding terms. This clashes with the concept of *polysemy*: a word can have multiple meanings depending on the context where it appears; word disambiguation is a research problem in computational linguistics, Britton (1978) estimates at least 32% of words used in the English language are ambiguous.

Contextualized representations and language modeling

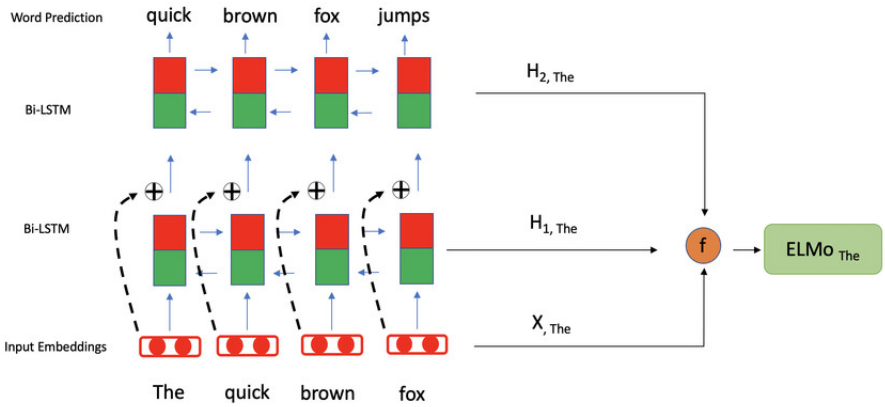

Figure 2. Visualization of ELMO word embeddings from Chiu and Baker (2020).

From bigrams we can expand language modeling to longer sequences, **n-grams**, and model the probability according to Equation 1. Recurrent neural networks (RNN), such as long-short term memories (LSTM) from Hochreiter and Schmidhuber (1997), can be employed to process word sequences: once defined a vocabulary, words are

identified with an id and projected to highly dimensional embedding vectors; the embeddings are fed in sequence to a RNN cell that updates an internal state, which at the end of the sequence will encode the sentence context. Peters et al. (2018) builds upon RNNs an architecture that processes text sequences forward and backwards (bi-directional LSTM) with stacked layers: a word embedding w_i is computed as the linear combination of the RNN cell hidden states at time i over the stack; the network is trained to perform language modeling and this approach is named **ELMO** (Embeddings from Language Models).

Recurrent neural networks present limitations for long sequences: a fixed-size vector cannot fully encode representations of long bodies of text; moreover, training is slow as words are processed in sequence and long sequences affect the quality of learning ("vanishing gradients"), as shown by Pascanu, Mikolov, and Bengio (2012). More recently, a novel architecture allowing processing sequence has replaced RNNs: transformers, introduced by Vaswani et al. (2017). Attention is the core computational block: each word embedding is projected into a query, a key, a value and matched to any other word in the input sequence via query-key similarity; the output sequence is represented by the word values weighted by the word by word similarities. This simple approach overcomes the problem of modeling long sequence dependencies, Dosovitskiy et al. (2020) showed this successfully applies to visual signals as well.

In correspondence to ELMO, Bidirectional Transformers in Language or **BERT**, Devlin, Chang, Lee, and Toutanova (2018), have been proposed to understand and encode language sequences using the above mentioned attention blocks. The training objective is **masked language modeling**: predicting the missing words in a sentence; words are initially converted to tokens ("word-pieces") and projected to embeddings as in RNNs.

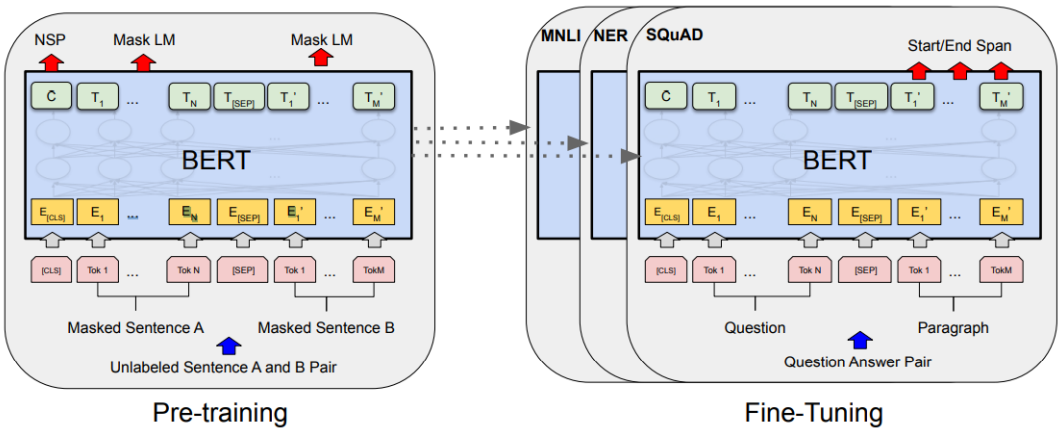

Figure 3. BERT pre-training with masked language modeling and task specific fine-tuning, Devlin et al. (2018).

To generate language sentences such as answers, an auto-regressive variant of transformer is used: Generative Pre-training Transformers or **GPT**, introduced by Radford, Narasimhan, Salimans, and Sutskever (2018). As in RNNs, the training objective consists in predicting the next most likely word occurring in a sequence, an attention mask is used for the attention blocks to attend only the context tokens.

In the attempt to map LM embeddings to brain recordings in processing spoken or written words, Hollenstein, de la Torre, Langer, and Zhang (2019) and Caucheteux and King (2022) report better matching for transformer-based architectures over RNNs and word2vec. I thus focus on studying BERT and GPT to answer the question: to what extent do artificial word representation resemble to brain activity? If present, what linguistic features rule this analogy?

Mapping linguistic representations to brain activity

A major question in neuroscience is how conceptual knowledge is encoded in the brain, how it is organized and how it constructs meaning. Neuroscientists rely on brain imaging techniques to capture mental states: functional magnetic resonance imaging (fMRI) allows to capture cerebral blood flows, which in high concentrations represent intense neural activity. Ishai, Ungerleider, Martin, Schouten, and Haxby (1999) and Haxby et al. (2001) show that spatial patterns in fMRI images can be matched with visual inputs such as objects and faces. In researching representations for the meaning

of words, psychologists focused on lists of terms associated by individuals; linguists associated semantic roles to verbs and nouns; alternatively, computational linguists claimed that word meaning is best captured by the distribution of co-occurring words. Mitchell et al. (2008) takes from the statistical definition of word meaning: given a trillion-words text corpus, nouns are represented by a vector of co-occurrence counts with respect to the finite set of semantic features (verbs):

```
"see", "hear", "listen", "taste", "smell", "eat", "touch", "rub", "lift",
"manipulate", "run", "push", "fill", "move", "ride", "say", "fear", "open",
"approach", "near", "enter", "drive", "wear", "break", "clean".

"bear" = {"see": 0.52, "hear": 0.52, "listen": 0.486, "taste": 0.214, ...}
```

.

A predictive model is proposed to learn fMRI patterns of nouns under two assumptions: semantic features of a word are reflected in its use in large bodies of text; the relationship between word vectors and brain activation patterns is assumed to be linear, which is consistent with the frequent use of linear models in analyzing fMRI activity, Friston (1995). Neural activity in the brain is approximated by a matrix of unitary locations named **voxels** v: in this study, the activation of an area y_v is modeled as the linear combination of semantic features $f_i(w)$ for an input word w, where the contribution c_{vi} of each feature i is learned through a regression problem:

$$y_v = \Sigma_i^n c_{vi} f_i(w) \tag{3}$$

fMRI data is collected over 9 subjects: 60 word-image pairs are presented to each subject for six repetitions (or epochs); as activity for over 20.000 voxels is recorded, the authors select the 500 most "stable" ones with the lowest variance over epochs. The model is trained on 58 samples and evaluated on the two held-out ones for all the possible combinations. To compute the accuracy, the model predicts voxel activity for the two unseen words: the obtained activations (e.g. p_1 for word w_1) shall correspond

by cosine similarity (*"cosim"*) to the ground truth (e.g. i_1) measurement rather than the other test sample (e.g. i_2), such that:

$$cosim(p1, i1) + cosim(p2, i2) > cosim(p1, i2) + cosim(p2, i1) \tag{4}$$

As a result, a mean accuracy of 0.77 over the 9 subjects is achieved. Cortical areas where activity has been predicted with the highest accuracy involve mainly the left inferior, temporal, motor cortex, inferior frontal, orbital frontal and occipital cortex; this is consistent with the general view that the left hemisphere is highly employed in semantic representation. Above-chance accuracies are obtained also by testing the model to predict the activity for words that belong to semantic categories excluded at training time (e.g. animals, mean accuracy: 0.70), or semantically similar words whose fMRI activity can be difficult to distinguish (mean accuracy: 0.62). Eventually, the distribution of learned weights for specific semantic features over the voxel space reveal analogies in line with theory (Figure 4): the fMRI image for "push" shows strong activations in the right postcentral gyrus, related to the coordination of complex movements; strong activations in the image for "eat" are found in the opercular cortex, which is assumed to be associated with the perception of taste.

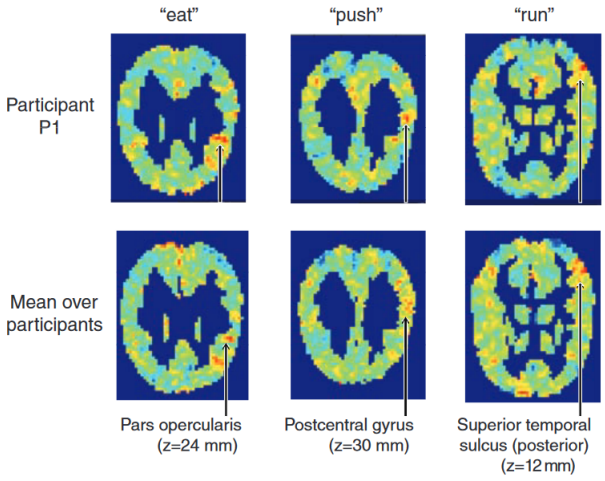

Figure 4. Spatial distribution of weights for semantic features "eat", "run", "push" from Mitchell et al. (2008).

This study sheds light on new ways to study brain activity: firstly, computational language representations can indeed be mapped to brain recordings with an acceptable

level of generalization; predicted fMRI images for the same words are consistent over multiple individuals, suggesting that brain responses to stimuli can arguably be general to groups of subjects; neural representations for nouns are partially grounded by sensory-motor features.

Basing on this approach, I have firstly reproduced the results obtained by Mitchell et al. (2008) and subsequently tested different computational word representations to predict the same fMRI images: non contextualized embeddings with Global Vectors for Word Representation or **GloVe**, from Pennington, Socher, and Manning (2014), based on matrix factorization and similar to word2vec; non contextualized BERT embeddings of one-word sequences; for context-based BERT embeddings, given 1.000 context sentences per word I consider the dominant meaning (the centroid of the largest cluster from K-means) and the average context vector, suggested by Chersoni, Santus, Huang, and Lenci (2021).

Embedding type	Accuracy (↑)
25 semantic features	$0,78 \pm 0,41$
GloVe	$0,71 \pm 0,46$
BERT (no context)	$0,68 \pm 0,47$
BERT (avg. context)	$0,55 \pm 0,44$
BERT (dominant meaning)	$0,71 \pm 0,45$
random vectors	$0,32 \pm 0,47$

Table 3. Mean accuracy scores for different word embeddings in predicting fMRI images for 2 held-out words.

These experiments suggest that computational word embeddings successfully map to fMRI images with accuracy above random level. I hypothesize the differences in performance could be caused by the source of grounding: Mitchell et al. (2008) pair pictures with words, while the only modality used to train GloVe or BERT is text; moreover, in this experiment contextualized embeddings are based on 1.000 randomly sampled paragraphs per word with no guarantees about bias. It is worth noticing that

the ground truth fMRI images represent single words without a context: this may justify why contextualized BERT embeddings perform worse than static GloVe vectors; a higher accuracy is achieved if the contextualized vector is reduced to the word's dominant meaning among 1.000 random context sentences.

To further understand word ambiguity, I visualize the meanings of a term given its representations from 1.000 sentences: contextualized BERT embeddings are clustered with K-means, with the optimal $k \in [0, 20]$ that maximizes the silhouette score (infra and intra cluster distance). Word representations are extracted from the last layer of a pre-trained BERT model, as it is supposed to learn abstract, semantic features. To extract the dominant meaning, the centroid of the largest cluster is considered, e.g. cluster 1 in the following example. Eventually, the labeled embeddings are projected to a bi-dimensional feature space with tSNE for visualization. Figure 5 illustrates the identification of 5 distinguished clusters of meaning for the word "bar", a known ambiguous term in the English language.

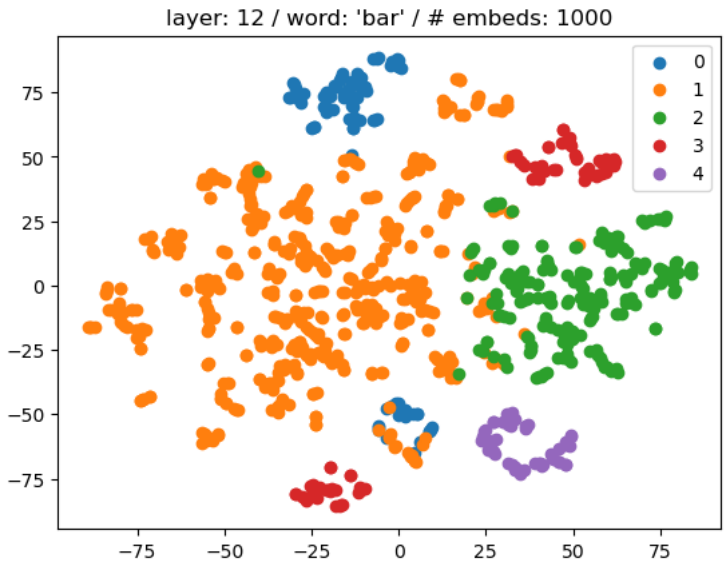

Figure 5. tSNE with d = 2 for 1,000 random context embeddings of the word "bar".

Predicting brain activity from context sentences with large language models

As anticipated, transformer-based networks motivated a paradigm shift in deep learning and natural language processing. Radford, Wu, et al. (2018) and Brown et al. (2020) showed that by scaling the model size and the amount of training data, the amount of tasks that can be solved and the relative accuracy increase. The underlying assumption is that the learned representations of language are general enough to detach from mere language patterns and move to the composition of abstract concepts.

Caucheteux and King (2022) scale up experiments on predicting brain activity in two ways: by adopting large language models for word representations to analyze their differences at varying depths of the network; with respect to existing literature, by collecting brain activity from a larger group of subjects (102) over sequences of 9-15 words. This work reports three main findings: language model activations linearly map to brain activity associated with reading; activations from the intermediate layers of transformers best map to brain activity; the quality of such mappings strongly depends on the level of the model's language abilities.

Initially, the authors test the hypothesis that neural responses to words and sentences are consistent across subjects. A linear regression model is trained to predict the fMRI response y to a word in a held-out subject from the average brain response to the same stimulus across the other individuals, namely the "template brain activity" x. A high Pearson R correlation with statistical significance (Wilcoxon two-sided test) between predicted and truth activity patterns confirm the presence of consistent neural responses to linguistic stimuli.

Consequently, similarly to Mitchell et al. (2008), a voxel-wise linear regression model is fitted to predict fMRI activity from computational word representations. Three types of inputs are tested: "visual" embeddings computed by a Convolutional Neural Network (CNN), Lecun and Bengio (1995), that recognizes characters from displayed words; non-contextualized "lexical" embeddings from the projection layer (token id to vector) at the beginning of a language model; contextualized "compositional" embeddings computed in the middle attention blocks ($\frac{1}{2}$th and $\frac{3}{4}$th) of the model.

The learned network to brain mappings provide supporting evidence on the hierarchy of neural activity in response to language, providing also hints on the factors determining the algorithm-brain analogy. Firstly, for all the types of representations the *brain score* (Pearson's R correlation between prediction and ground truth) surpasses chance-level, with peaks in different cortical regions depending on the type of input (Table 4).

Embedding type	Peaking area	R	Stat. Significance
visual	visual cortex (V1)	0.022 ± 0.003	$p < 10^{-11}$
lexical	left superior temporal gyrus	0.052 ± 0.004	$p < 10^{-13}$
lexical	inferior temporal cortex, middle frontal gyrus	0.053 ± 0.003	$p < 10^{-15}$
contextual	superior temporal gyrus	0.012 ± 0.001	$p < 10^{-16}$
contextual	angular gyrus	0.010 ± 0.001	$p < 10^{-16}$
contextual	infero-frontal cortex	0.016 ± 0.001	$p < 10^{-16}$
contextual	dorsolateral prefrontal cortex	0.012 ± 0.001	$p < 10^{-13}$

Table 4. Peaks of brain scores and statistical significance for word embeddings in cortical regions, Caucheteux and King (2022).

Middle layer activations sistematically outperform input and output layers, potentially indicating that representations that best resemble brain activity are to some extent abstract, but not necessarily they depend on the output layers of the network and thus on the training task.

With source-localized MEG, activations in the above mentioned cortical regions are tracked over time: from the presentation of the stimulus, different word representations best match neural activity at different steps in time, suggesting that information processing in the brain is distributed across multiple modalities and levels of abstraction. At 100ms, the brain scores for visual embeddings peak in V1, followed by word embeddings matching activity in the left posterior fusiform gyrus at around 200ms and the left temporal and frontal cortices at 400ms; compositional (deep) embeddings eventually map to activity in multiple bilateral regions at around 1000ms, Figure 6.

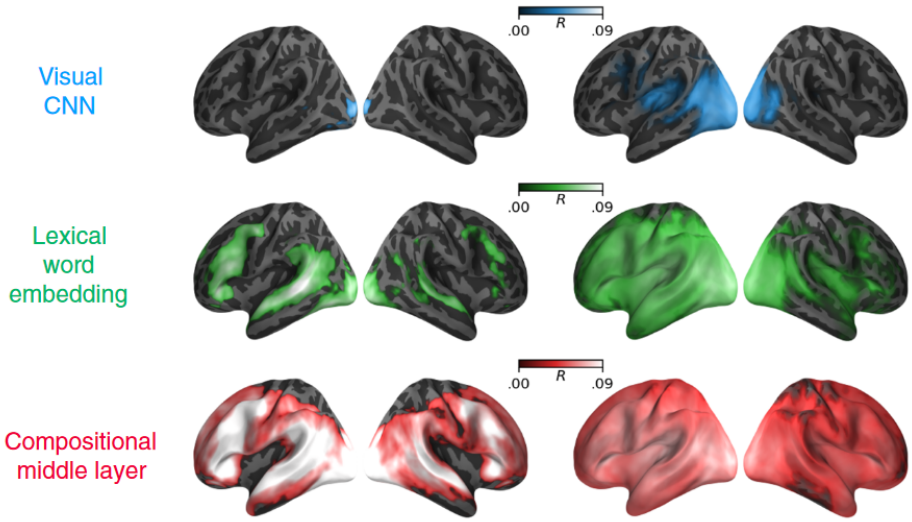

Figure 6. Localized brain scores for fMRI (left) and MEG (right) mappings of visual, lexical and compositional embeddings from Caucheteux and King (2022).

Eventually, language model performance in the training task is found to highly correlate with brain scores. The authors test 32 different architectures varying in depth and size of the embeddings. Early experiments with untrained models (randomly initialized weights) reveal that their activations can be mapped to brain activity consistently (Pearson's $R = 0.019 \pm 0.001$ score, $p < 10^{-16}$), which suggests that transformers partially map to brain activity independently of their acquired language abilities. In models trained to perform language modeling, brain scores strongly correlate with top-1 accuracy (both for causal or masked language modeling, that is predicting the next or a missing word), with the highest brain scores achieved by middle layers ($R = 0.81 \pm 0.02$), followed by output layers ($R = 0.63 \pm 0.03$) and input layers ($R = 0.39 \pm 0.03$). However, the networks with the best language performance do not yield the highest brain scores: arguably, this is potentially caused by overfitting to the training task, which does not capture more complex dynamics of language generation in the brain, including long-range and hierachical dependencies. Moreover, transformers differ from their biological equivalent as information flows in one direction (without recurrent paths) and significantly larger amounts of examples are required to learn the same ability.

From this study I derive two further questions to investigate: to test whether different learning rules can improve algorithmic convergence to brains in language processing; to verify if contextualized embeddings may hold representations reflected in cortical areas dedicated to abstract reasoning.

Abstraction I: modelling brain representations of abstract concepts

Kaiser, Jacobs, and Cichy (2022) focus on abstract concepts to investigate the analogies between representations in computational models and the human brain. In their study, 19 participants are asked to elaborate 10 stories each, by using 61 nouns displayed in different orders of sequence given an initial background context. Brain activations are collected through fMRI as every noun is prompted to the subject; computational representations for the same nouns are extracted from a word2vec model trained on the SdeWaC corpus, Faass and Eckart (2013), containing 45 million German sentences. With searchlight analysis, a representational dissimilarity matrix (RDM) is computed on localized neural activities recorded when the 61 nouns are displayed; similarly, a RDM for the same set of nouns is computed using the last hidden layer activations extracted from a word2vec model.

Without employing models that explicitly learn a "network to brain" mapping, as in the previously mentioned works, the authors identify cortical areas that are assumed to be involved in abstract knowledge processing; these regions are also matched with representations from computational models of abstract concepts and their compositions. Firstly, high correlation is found between the network and brain RDMs in the left interior parietal cortex (IPC), including the angular gyrus, the superior parietal and middle occipital cortices and part of the right IPC. Arguably, this suggests that IPC plays a crucial role in concept coding, and the structure of its representations is partially reflected in the vector space constructed by computational models. However, analogies in the structure of the two representational spaces may be also explained by the linguistic properties of nouns, as word2vec models organize terms basing on their co-occurrence in written language. Finally, the task introduced in this study highlights

the role of the angular gyrus in the dynamic use of abstract knowledge, compatibly to Davis and Yee (2018) and Price, Bonner, Peelle, and Grossman (2015); this does not exclude a possible relevance for the angular gyrus in processing also concrete nouns, comparing the representations of the two types requires further study.

High correlations of the RDMs in IPC are maintained when the model is trained only on abstract ($R = 0.36, p = 10^{-2}$) or concrete ($R = 0.73, p = 10^{-3}$) nouns. Moreover, brain scores are directly proportional to the size of the training set, with a lower bound of 100,000 samples (0.02% of the dataset) estimated to obtain above-chance positive correlation. This suggests that the representational structures of both concrete and abstract nouns overlap with the organization of abstract knowledge, which is supposed to emerge through compositionality.

Liao, Chen, and Du (2023) specifically test the capabilities of transformer models in hierarchically organising abstract and concrete nouns. With WordNet, a dataset of hypernyms (e.g. "furniture" is a hypernym for "bed") is constructred: each sample is a *noun-noun* pair, the model is prompted to detect whether the relationship two nouns is hypernymy or not. All the tested language models can detect hypernymy with above-chance accuracy, however the performance is consistently worse in abstract nouns with respect to concrete ones.

Model	Concept type	F1
BERT	abstract	0.8435
BERT	concrete	0.8912
T5	abstract	0.8543
T5	concrete	0.8888
ChatGPT	abstract	0.4410
ChatGPT	concrete	0.7304

Table 5. F1 scores in hypernymy detection with large language models, Liao et al. (2023).

These studies suggest that language models, with and without context, can

understand the structure of abstract knowledge but not as extensively as simple nouns: terms can assume a symbolic valence with semantic relationships with other concepts; as in Santoro, Lampinen, Mathewson, Lillicrap, and Raposo (2022), I argue that language is insufficient to fully grasp their meaning and multimodal grounding is necessary (vision, audio, interaction). In the following sections I will also discuss the abilities and limitations of neural networks in compositionality, which underlies abstract reasoning in the hierachical processing hypothesis.

Abstraction II: predicting semantic comprehension from brain activity

To what extent do language models understand concepts? Caucheteux, Gramfort, and King (2022) investigate how GPT-2 brain scores (Pearson's R correlation) vary with semantic comprehension of short stories: 101 subjects listen to 7 narratives and compile a questionnaire at the end of each one, fMRI activity is recorded during the process. Activations from different layers of GPT-2, Radford, Wu, et al. (2018), are used to predict voxel activity in the form of fMRI images: brain scores are eventually correlated with semantic scores from each story's survey. A high level of understanding of the narrative is linked to better network-brain mapping, which may also be influenced by the processing of speech, lexical and conceptual features in the input language.

Across all voxels, the correlation between brain and semantic scores reaches $R = 0.50, p < 10^{-15}$ with peaks in bilateral temporal, parietal and prefrontal cortices. fMRI activity is predicted also with low-level audio features as word representations (word rate, phoneme rate, stress, tone): the correlation between brain and semantic scores is positive ($R = 0.17, p < 10^{-2}$) but weaker with respect to GPT-2 predictions ($\Delta R = 0.32$); low-level activations best predict activity in the left superior temporal cortex. This analysis may suggest that audio processing is another indicator for semantic comprehension: speech understanding is linked to attention and it reflects in the acquired semantics.

Activations from the 8th layer in GPT-2 yield brain scores that best map to semantic comprehension in comparison with the word embedding layer (the first) and

middle layers. In deep learning literature, the level of depth at which a representation is computed in a network reflects the level of abstraction of information. Activations from the word embedding layer are considered to encode lexical features of language and best match activations in the superior-temporal lobe and in pars triangularis. Conversely, high-level representations from GPT-2 eighth layer encode contextual features and best map to voxels in superior-frontal, posterior superior-temporal gyrus and in both the triangular and opercular parts of the inferior frontal gyrus, known to be involved in high-level language processing.

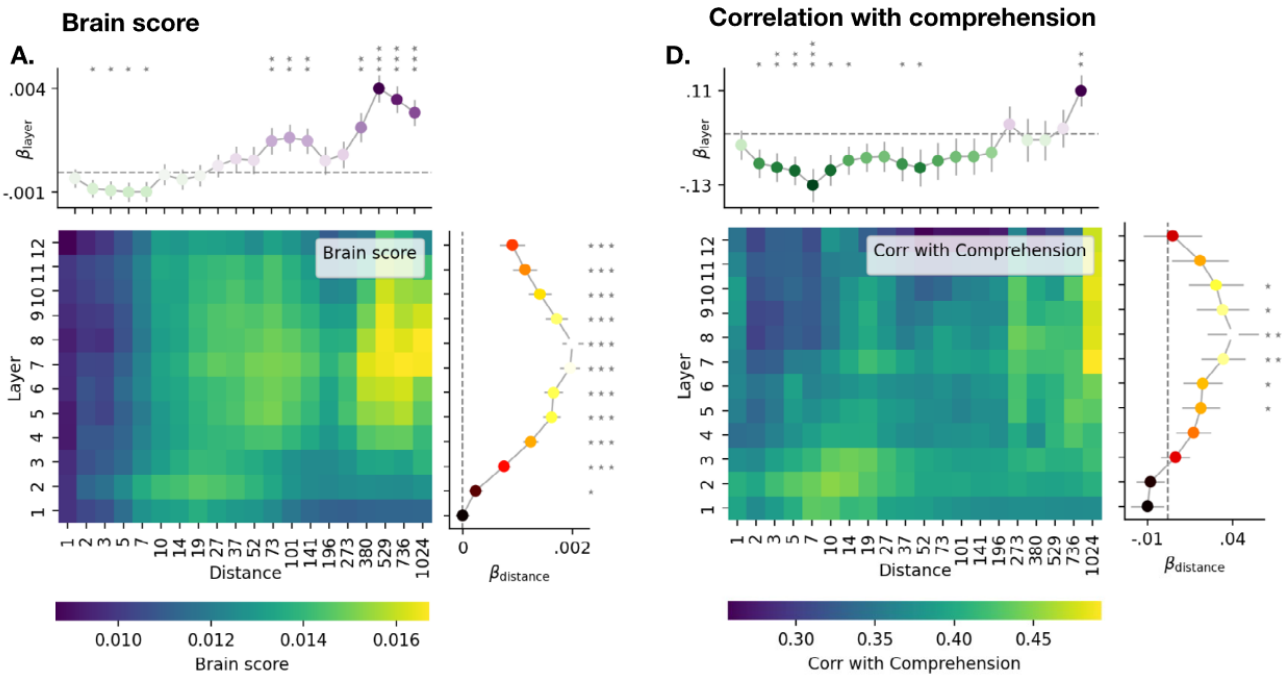

Figure 7. (A) brain and (D) semantic scores for a configuration (k, d), positive $\beta_{\text{distance}}, \beta_{\text{layer}}$ indicates scores are influenced by distance and depth respectively, Caucheteux et al. (2022).

Finally, a fine-level analysis reveals the relationship between the network depth, the span of attention to language and brain-semantic scores. By varying the depth of the layer k employed for word representations and the length d of the input sentence, the following relationships are found: both brain and semantic scores improve with increased attention span d (Figure 7); reduced attention spans improve scores for middle layers activations, suggesting that transformers may be more similar to the brain when

the attention span is increased as function of the depth; finally, brain regions associated to high-level concept processing are best predicted by activations from the deepest layers and brain-semantic scores improve with longer attention spans; areas accounting for low-level audio processing in the brain are best predicted by activations from the shallowest layer of the network and scores are not influenced by the length of the inputs.

This study provides evidence supporting the idea of hierarchical language processing in transformer models and how contextual information can improve abstract reasoning. Computed representations at varying depths map to different areas of the brain, finding evidence of reasoning at different levels of abstraction.

Hypotheses for language models to converge with brain activity

Decoding computational representations with semantic features

I have discussed various methods proposed to encode word semantics in sparse or dense representations. With reference to Mitchell et al. (2008) and Caucheteux and King (2022), I argued that engineered or learned word embeddings can predict brain activity patterns, which are assumed to be general across multiple subjects: under these assumptions, Caucheteux et al. (2022) investigated how semantic comprehension and audio processing modulate these network-brain mappings and found positive correlation. Eventually, Kaiser et al. (2022) observed analogies in the structure of abstract and concrete concepts between artificial and biological neural systems. Assuming that thus word embeddings encode in a structured way concepts with a certain degree of understanding, I overview different techniques used in literature to test these hypotheses.

Word embeddings are computed with what have been generally defined as Distributional Semantic Models (DSMs), Lenci (2018), ranging from count-based to predictive methods. However, the representations in these models are not interpretable: content is encoded in learned numerical representations across multiple dimensions; vectors are a "holistic" representation of knowledge and word meaning can be inferred by computing the relative position with respect to other concepts in the

representational space. Alternatively, probing tasks, Ettinger, Elgohary, and Resnik (2016), have been commonly used in literature to look for subsets of embeddings that represent specific semantic features (e.g. positive sentiment words): regression models or neural networks are employed as classifiers for this task; however, this technique does not address the problem of explainability.

Chersoni et al. (2021) propose an alternative approach taken from computational neuroscience: research in neurosemantic decoding attempts to identify mental states represented by recorded brain activity, usually through functional magnetic resonance imaging (fMRI). More practically, fMRI images of words are linearly mapped to human similarity scores for different semantic features. With respect to neurosemantic decoding, the authors don't consider word embeddings as the prediction target but rather as the input space to project into interpretable features; the goal is thus to understand what semantics are best encoded in language models, how they impact performance in semantic probing tasks and whether they can explain the difference between contextualized (e.g. BERT, ELMO) and "static" word representations (e.g. word2vec, GloVe).

Embeddings from a set of predict (**ELMo**, Chiu and Baker (2020); **BERT**, Devlin et al. (2018); **SNGS**, Mikolov et al. (2013); **FastText**, Bojanowski, Grave, Joulin, and Mikolov (2017)) and count models (**GloVe**, Pennington et al. (2014), **PPMI**, Levy and Goldberg (2014)) are compared to find the best mapping to semantic feature vectors. For models that encode context, a word is represented by the average embedding over 1,000 sentences randomly sampled from Wikipedia, reflecting the hypothesis that context-independent representations are an abstraction of token exemplar concepts, Yee and Thompson-Schill (2016). Semantic features from Binder et al. (2016) are fixed as prediction target: they consist in valence ratings in a 0-6 scale for 65 properties (noun, verb, adjective); this dataset provides thus semantic ratings for 534 abstract and concrete words; empirical evidence in neurocognitive research reports the relevance of these features in conceptual organization, Chersoni et al. (2021). To map LM activations to Binder features, a PLS regression model is used with different values

for K components to be tested, $K \in \{30, 50\}$ is found to be optimal. The models are evaluated with leave-one-out cross-validation, by computing the Mean Squared Error and the top 1-5-10 accuracy to predict the left out word among the most similar feature vectors. As baseline, a PLS regression model is trained to project randomly generated word embeddings.

Model	(↑) Top-1	(↑) Top-5	(↑) Top-10	(↓) MSE
PPMI.w2	0.14	0.42	0.57	0.16
PPMI.synf	0.14	0.46	0.61	0.15
PPMI.synt	0.10	0.36	0.54	0.16
GloVe	0.18	0.43	0.58	0.16
SGNS.w2	0.19	0.49	0.64	0.15
SGNS.synf	0.20	0.55	0.71	**0.14**
SGNS.synt	0.23	0.57	0.74	**0.14**
FastText	0.20	0.53	0.70	**0.14**
ELMo	0.22	0.50	0.68	0.16
BERT	**0.30**	**0.59**	**0.76**	0.15
Random	0.00	0.01	0.01	0.30

Table 5. Evaluation scores for LM-semantic features mapping with PLS Regression, K=30, Chersoni et al. (2021).

The results confirm the superiority of contextualized embeddings over static DSMs, as demonstrated also by Vulić, Ponti, Litschko, Glavaš, and Korhonen (2020), followed by the syntactically enriched versions of skip-gram vectors (SNGS), Table 5.

A further interesting analysis consists in computing the Spearman correlation coefficient between the predicted and the ground truth semantic feature vectors. Strikingly, as shown in Figure (7), the highest correlation is achieved for features belonging to the cognitive, causal and social domain, for which psycholinguistic studies theorize that language, and thus text, is the main source for this type of information, Vigliocco and Gareth (2007). Lower correlation corresponds to somatosensory features

of concrete concepts; this occurs also for spatial and temporal features, compatibly to the hypothesis that temporal concepts are based on spatial references and language grounded with other modalities (e.g. vision or social interaction) is fundamental, as claimed by Santoro et al. (2022) and Binder et al. (2016).

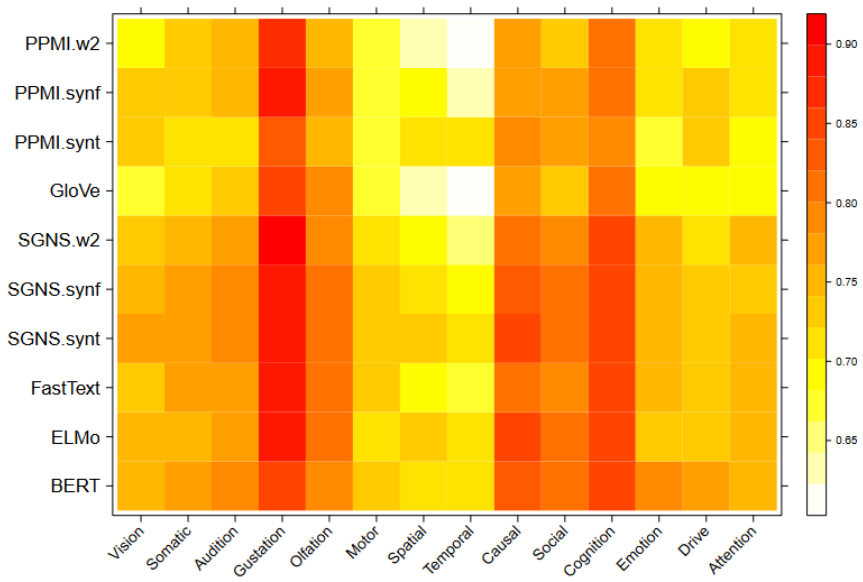

Figure 7. Average Spearman correlation per semantic domain between the predicted and the original Binder ratings, Chersoni et al. (2021).

An ablation study to query semantic concepts driving brain mapping

To further understand the overlappings between language models and the human brain, I conduct an ablation study to assess the influence of different semantic features on brain mapping. Similar to the work of Chersoni et al. (2021), a multilayer perceptron network with 100 hidden units is trained to map GPT-2 word embeddings to word fMRI images from the narrative Harry Potter dataset, Wehbe, Vaswani, Knight, and Mitchell (2014). For computational efficiency I use the post-processed fMRI images of 500 random voxels provided by Hollenstein et al. (2019). This study differs in the adopted methodologies with respect to Mitchell et al. (2008) for two aspects: firstly, brain activity is collected from subjects reading words in a continuous a story, thus activations are contextualized. Secondly, compatibly to Utsumi (2020) and Chersoni et al. (2021), I assume that voxels can condition each other and activate due to spatial or

topological proximity: a multivariate multiple linear regression model is chosen over independent voxel regressors.

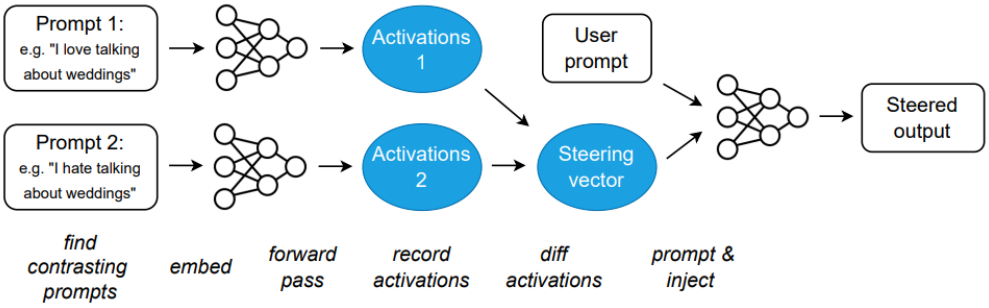

Figure 8. LM activation steering by computing a semantic vector from vector differences, Turner et al. (2023).

To extract contextualized embeddings from GPT-2, I consider the intermediate representations from the 8th layer of the network, as shown by Caucheteux, Gramfort, and King (2023) to best map to brain activity. In line with the methodology from the same authors, each word is concatenated to a context window of 8 preceeding words (= 40 GPT tokens), corresponding to 3.15 seconds of speech with 2.54 words/second.

The influence of different semantic features is measured by subtracting them from each word's context. To achieve this, I employ a recent alignment technique named **activation steering**, introduced by Turner et al. (2023): given a set of terms representing a concept, e.g. "man", "woman", female", "male" for *gender*, the average GPT embeddings from layer K represent the semantic category to remove; by subtracting this vector from activations at layer K at inference time, I expect to obtain an output that is deviated off the subtracted direction. In Figure 8, Turner et al. (2023) use the difference from a "love" and a "hate" sentence to add a "love activation" and generate positive sentences with GPT-2. Similarly, I compute the steering vector for three semantic categories: *time*, *gender* and *sensory-motor* (taken from Mitchell et al. (2008)).

```
sensory_motor = ["see", "hear", "listen", "taste", "smell", "eat",
"touch", "rub", "lift", "manipulate", "run", "push", "fill",
```

```
"move", "ride", "say", "fear", "open", "approach", "near",
"enter", "drive", "wear", "break", "clean"]

gender = ["male", "female", "boy", "girl", "man", "woman"]

time = ["when", "yesterday", "ago", "last week", "last month",
"last year", "still", "yet", "while", "when", "soon", "then",
"next week", "next month", "next year", "tomorrow",
"the day after tomorrow"]
```

The predictive model is trained and evaluated for each ablation with cross validation on 20 splits (average), using the metrics introduced by Chersoni et al. (2021): mean squared error (*MSE*); top-N-accuracy (*Acc@Top5*, N = 5), that is the fraction of times the ground truth vector is among the closest N to the prediction. The scores for different ablations (removed semantic categories) are compared with a baseline consisting in GPT-2 without any steering; the difference in accuracy between each model and the baseline is reported as "Δ". I use a two-tailed T-test to assess whether the difference between each model score and the baseline over 20 splits is neglibile (null hypothesis).

The reported results firstly suggest that GPT-2 embeddings linearly map to brain activity from contextualized words, as shown by Caucheteux and King (2022), Caucheteux et al. (2023). Moreover, the relevance of features from different semantic categories emerges: removing information about gender, sensory-motor actions and time affects brain mapping with high statistical significance, Table 6. I hypothesize the importance of each feature depends on its presence in the read story: information about the subjects such as gender, is critical for overall understanding; sensory-motor and temporal features provide further context.

Removed semantic category	MSE (↓)	Acc@Top5 (↑)	Δ	p-value
None (baseline)	$0,0148 \pm 0,0159$	$0,9838 \pm 0,0309$		
Sensory-motor	$0,0722 \pm 0,0235$	$0,7178 \pm 0,2033$	$-0,2660$	$3,3623 \cdot 10^{-10}$
Temporal	$0,0735 \pm 0,0141$	$0,6843 \pm 0,1437$	$-0,2995$	$2,1118 \cdot 10^{-14}$
Gender	$0,0814 \pm 0,0225$	$0,5902 \pm 0,1905$	$-0,3936$	$2,9526 \cdot 10^{-12}$

Table 6. Brain mapping scores for models steered away from different semantic features. Δ and p-value are computed with respect to the baseline.

From this digression I derive further hypotheses: a narrative communicated through language yields knowledge on a set of entities with different levels of abstraction (concrete objects vs. symbols); assuming semantic comprehension is centered on a subset of subjects, features related to such entities are crucial for overall understanding and they are reflected in brain activity.

Evidence of predictive coding theory in the brain

According to the hierarchical predictive coding hypothesis, the brain continuously anticipates sensory inputs with predictions that originate top-down in different regions, each area learns and refines an internal model of the environment based on the statistics of inputs. The difference between predictions and inputs, the prediction error, is propagated to areas dedicated to high level processing to update the internal models. In testing these assumptions, Wacongne et al. (2011) looks for evidence of two phenomena: responses to stimuli of different semantic levels should differ in the activation pattern, suggesting hierarchical processing; brain responses should reflect predictive behavior rather than passive adaptation to the upcoming stimuli, suggesting an active role in anticipating them.

In this study, a group of subjects is presented auditory sequences of tones organized in three possible blocks: the "xxxxx" block, consisting in a starting sequence of 5 equal tones repeated 25 times, then followed by 100 repetitions of random sequences that can be either the same (75%, "local standard"), differing by the last tone (15%, "local deviant") or only 4 equal tones (10%, "omission"); the two other blocks

follow the same paradigm but the starting and most likely sequences are "local deviant" and "omission", respectively. Analyzing responses to sequences with omission is crucial: depending on the preceeding pattern, "xxxxx" or "xxxxY", evoked potentials similar to the missing stimulus are indicators of predictive behavior. The analysis is conducted separately with different brain imaging techniques: electroencephalography (EEG), magnetoencephalography (MEGm) magnetometers and MEG gradiometers (MEGg).

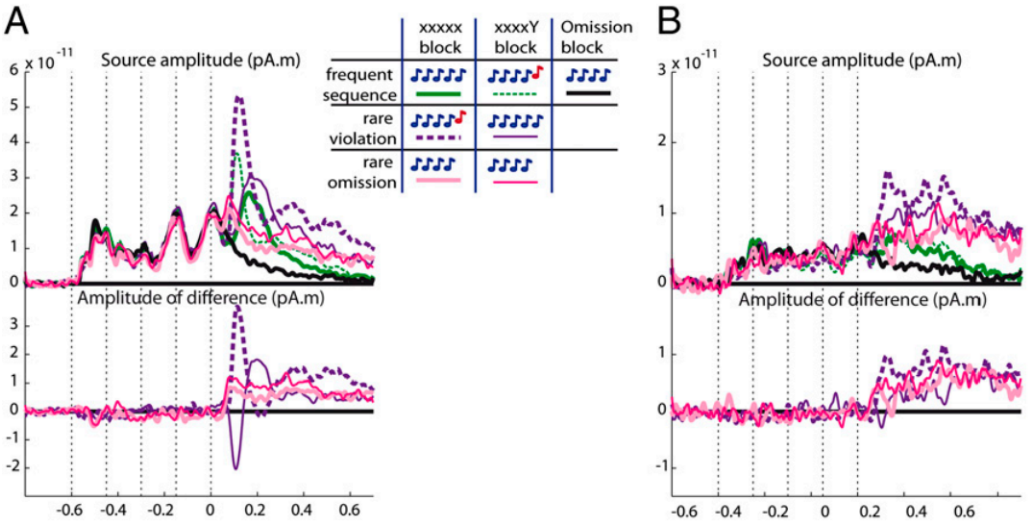

Figure 9. Activity in the auditory (A) and right precentral cortex (B) for the different combination of sequences and variations, Wacongne et al. (2011).

The effect of local deviance is reported in all the recordings: responses peak at 120ms after the onset of the fifth tone of the sequence; for both the local standard and local deviant blocks, the novelty response matches an activity pattern known in literature as "mismatch negativity" (MMN), which in this study is localized mainly in the temporal cortex. MMN response is observed to be "blind to the global deviant": when the starting sequence is switched to "xxxxY", this response pattern persists with a lower intensity, showing a lack of sensibility to the switch in the overall sequence and rather reflecting surprise to unexpected tone-by-tone transitions, Figure 9. Conversely, a later activity pattern is found to be distributed across the cortex, including particularly the prefrontal and parietal areas: this is defined as the "P3b wave", Wacongne et al. (2011), and it represents the "violation of violation", i.e. a monotonic "xxxxx" sequence following "xxxxY" doesn't meet the expectation; this activity pattern

is associated with brain processing that extracts and detects more abstract rules at sequence level, contrarily to tone-level changes to which MMN responses correspond.

The analysis of responses to omitted tone sequences conducted by Wacongne et al. (2011) thus support the idea of active prediction underlying language processing, reflected in recorded brain activity. Furthermore, the difference in the activation patterns between MMN and P3b highlights two levels of reactions to unexpected signals: predictions in the brain occur at multiple levels of abstraction, thus activity representing errors propagates across different cortical areas possibly at different time frames.

Further evidence is found in the experiments conducted by Heilbron, Armeni, Schoffelen, Hagoort, and de Lange (2022), which focus on the task of linking brain responses to unexpectedness of words. GPT-2 is used to extract semantic word embeddings and level of surprisal for the next words in sequence, brain activity is recorded with EEG and MEG.

Firstly, it is confirmed that continuous word embeddings better predict brain responses over representations based on single-word audio features; adding per-word unexpectedness (conditioned probability by context) as regression input is found to modulate predicted neural activity, which further suggests that prediction underlies language processing and errors are reflected in brain responses. Furthermore, the authors derive for each word different levels of features: syntactic, with part of speech tagging; phonetic, with a phoneme classification model; semantic, with the averaged GloVe embedding of the context words. The effect of each addition is measured by comparing the ground truth and the predicted next word: each feature improves brain mapping scores in different, unique areas, Figure 10; surprisal on the syntactic level modulates activity in focal and temporal areas; error on the semantic level explains later responses in a distribute set of cortical areas (including frontal).

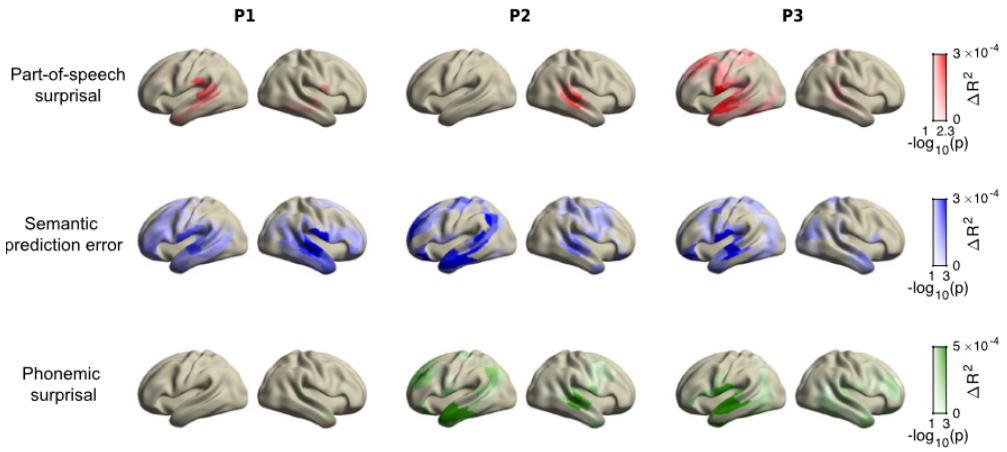

Figure 10. Unique patterns of explained variance in brain activity by lexical (POS), semantic and phonemic features, Heilbron et al. (2022).

Compatibly to Wacongne et al. (2011), this study suggests that while understanding speech, the brain performs prediction on different levels of abstraction. Consequently, I consider hierarchical predictive coding as a theoretical base to find analogies in language models: what linguistic features modulate similarity with activity in higher cortical areas; what features can improve convergence between algorithms and the brain.

Evidence of predictive coding theory from language models

The beginning of deep learning, arguably marked by the first applications in computer vision such as AlexNet, Krizhevsky, Sutskever, and Hinton (2012), determined a paradigm shift in designing neural networks: stacked layers of computational units can learn representations of data at varying levels of abstraction, allowing for complex tasks such as object classification, pixel-wise semantic segmentation (i.e. determining contiguous areas of the image that belong to an entity) or captioning. In natural language processing, transformers lead to modeling language distributions from internet-scale datasets thank to their capability to scale in number of layers, and thus algorithmic parameters by design, Brown et al. (2020). Kaplan et al. (2020) and Wei et al. (2022) examined the performance of language models of increasing size and found linear correlation with the amount of solved tasks and performance; this further

suggests that going deeper in neural architectures can improve their abilities in predicting abstract concepts as well as with lower level information.

In the attempt to reconcile hierarchical predictive coding theory with algorithmic language processing, Caucheteux et al. (2023) analyze computational and brain activations elicited by the processing of text stories. The experimental setup includes 304 individuals, each listening on average to 26 minutes of narratives, while neural activity is recorded with fMRI. GPT-2 XL, Radford, Wu, et al. (2018), is used to extract each word's representation conditioned by the preceeding context. Compatibly to Caucheteux and King (2022), activations from the 8th layer of GPT-2 best predict recorded fMRI, achieving the highest brain scores in voxels in the auditory cortex, the anterior and superior temporal areas.

As language models are trained to predict the next word in sequence, word representations are enriched in two ways to test predictive coding theory: firstly, next word embeddings are concatenated with the activations of $w = 7$ future words at distance d (position of the last word of the sequence with respect to the current predicted one), this improves brain scores by 23% ($\pm$9% across individuals) with $d^* = 8$. With "**forecast score**" $\mathcal{F}$, the authors refer to the brain prediction performance obtained by including future word embeddings. Secondly, GPT-2 is fine-tuned to predict not only the next word, but also the embedding vector of the future word at distance $d = 8$: this improves the accuracy by $\sim 2\%$ in predicting brain responses from frontoparietal areas, while no significant changes are found in auditory regions. These results suggest that predicting longer context windows leads to representations that better reflect brain activity, also when the model does not necessarily predict words: the authors argue that processing continuous representations, such as direct embedding vectors, lowers prediction errors that can occur due to the lack of concept of geometrical proximity with raw words.

Do different areas of the brain predict the same time window? Varying distances d^* are computed to maximize the forecast score $\mathcal{F}$ in each cortical region: long distances, $d^* > 9$, are matched with the inferior temporal gyrus (IFG), with respect to

the anterior superior temporal sulcus (aSTS), with $\Delta d = 0.9 \pm 0.2$. Forecast scores are also modulated by the depth of the GPT-2 layer k from which the activations are extracted: embeddings with $k^* > 6$ improve $\mathcal{F}$ in the association cortex, while activations with depth $k^* < 6$ improve $\mathcal{F}$ in aSTFS and Heschl's gyri ($\Delta k* = 2.5 \pm 0.3$), Figure 11. In conclusion, predictions of activity in frontoparietal cortices are found to depend on inputs with long contexts, with respect to areas linked to the processing of lower level features, e.g. audio.

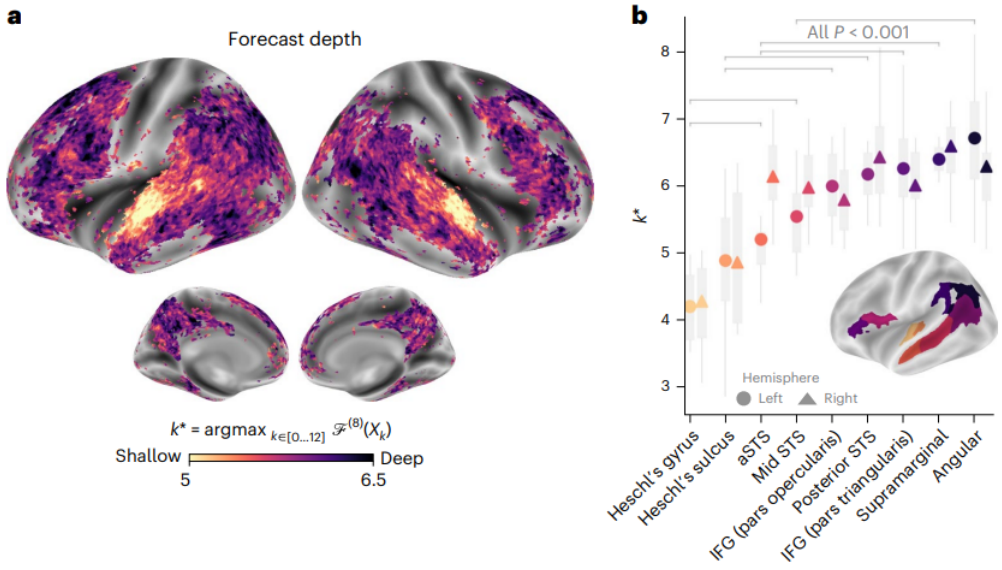

Figure 11. Depth k^* maximizing forecast score $\mathcal{F}$ per voxel(a) or brain region(b), Caucheteux et al. (2023).

As in Caucheteux, Gramfort, and King (2021), word embeddings are deconstructed into syntatic and semantic vectors to further investigate concept representations on multiple levels of abstraction. Syntactic vectors are computed by generating 10 possible sentences given the current context and then averaging the embeddings, X_{syn}; by construction, semantic vectors are computed by subtracting the syntactic component to the original word embedding, $X_{sem} = X - X_{syn}$. Forecast scores $\mathcal{F}$ significantly improve with semantic vectors only if concatenated with embeddings from long distance words (d = 8), peaking in the frontal and parietal lobes; conversely, syntactic embeddings maximize brain scores when added to words of relatively short distance (d $\leq$ 5), best mapping to superior temporal and left frontal areas. Overall,

semantic embeddings with long-range contexts map to activity in regions accounting for high-level processing: lateral, dorsolateral, inferior frontal and supramarginal gyrus; syntactic embeddings with short term context best predict low-level processing in the superior temporal sulcus and gyrus.

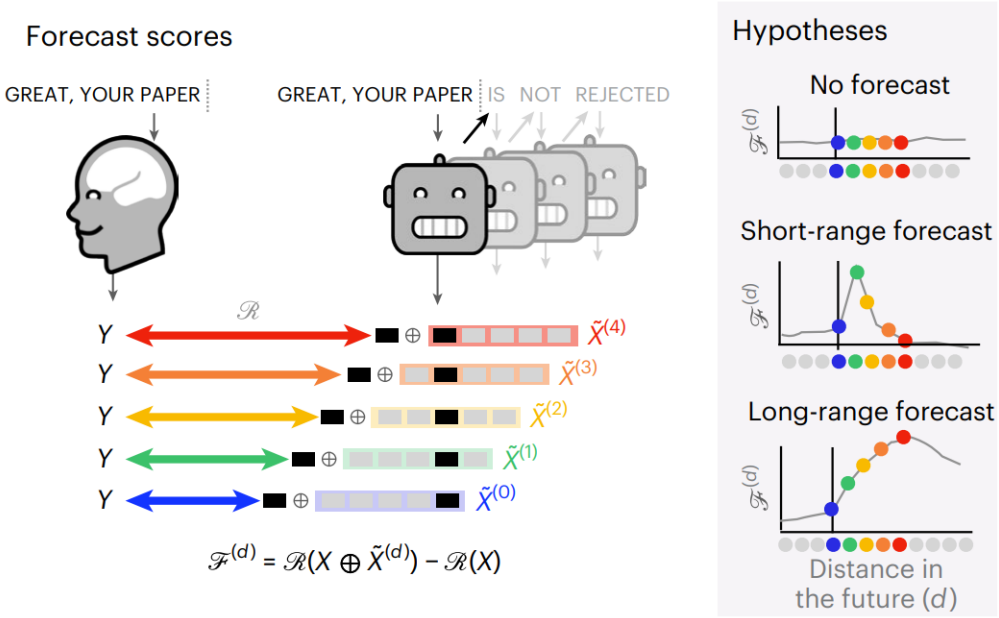

Figure 12. Forecast scores $\mathcal{F}$ are modulated by future words distance d and type of embedding (short range, long range), Caucheteux et al. (2023).

This study is found to be compatible with the hypotheses from Wacongne et al. (2011): by knowing the nature of computational representations, it is possible to verify correspondence to brain activity in processing more or less abstract concepts; evidence suggests that language models can generate predictions spanning across abstraction levels and contextual spans ranges, which are reflected in cortical areas theorized to account for high-level reasoning.

Challenging predictive coding theory as base of brain's processing

The studies from Chersoni et al. (2021) and Caucheteux and King (2022) suggest that next-word performance is positively correlated with the ability to predict neural activity from LM embeddings. From the perspective of predictive coding theory, which sees the ability to predict future events as the base of brain processing, this correlation

is tendentially interpreted as causation: Antonello and Huth (2023) challenge this vision by assessing a variety of linguistic abilities that condition brain mapping performance.

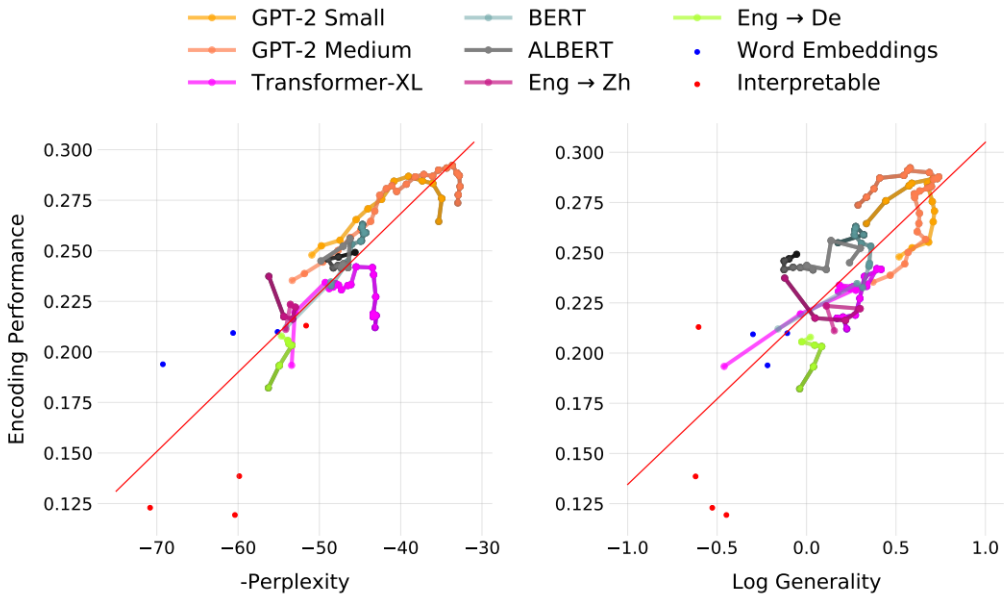

Figure 13. Correlation between brain mapping (encoding) performance and next-word prediction (perplexity), Antonello and Huth (2023).

In the first experiment, 97 different representations per word are extracted from contextual, non-contextual and seq2seq-translation language models (including BERT, GloVe, Transformer-XL, GPT-2). Correspondingly to each model, each voxel activity is modeled with linear regression, similarly to Chersoni et al. (2021) *encoding models*. Brain mapping performance is measured with Pearson's R correlation between the predicted and the actual activations on the held-out test set; next-word performance is measured by computing cross entropy over the distribution of predicted word embedding from the previous one in sequence with linear regression. An additional metric is introduced, the *representational generality performance* (RGP), defined as the average regression score for a (model, word) pair in predicting the word's representations from the other 96 models: this metric should represent to which extent a model learns linguistic features that well transfer regardless the training task (causal LM or masked LM) and context (GloVe or GPT).

Positive correlation between encoding models and language modeling performance is found ($R = 0.847$), in line with previous works; similarly, representational generality performance positively correlates with the performance of encoding models ($R = 0.864$). This suggests an alternative hypothesis to predictive coding: LM activations well map to brain activity as they encode generally useful representations independent from the training task; consequently, the causality argument involving the ability to predict the next word should be considered with caution.

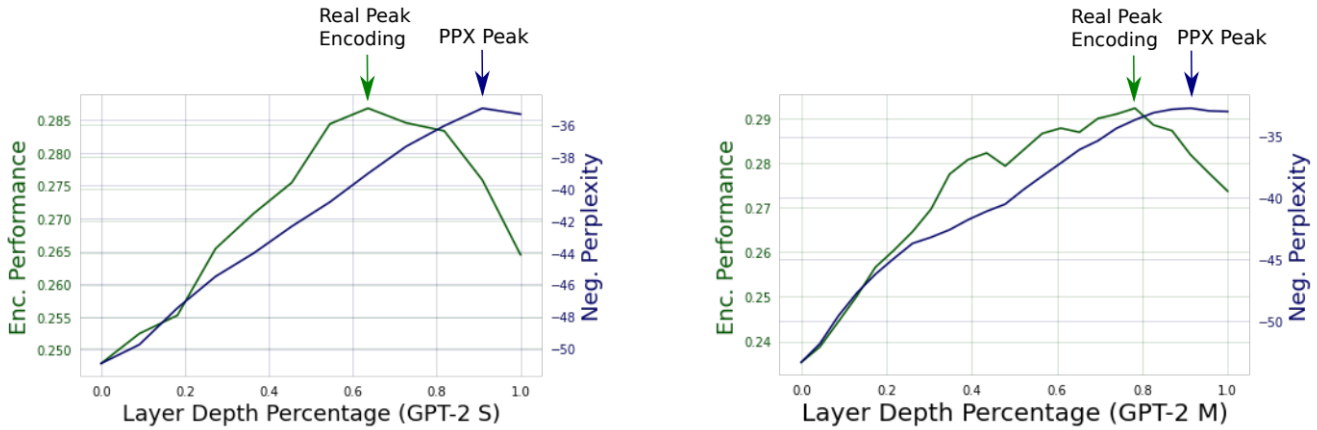

Figure 14. Negative perplexity and brain mapping (encoding) performance as a function of layer depth in GPT2-S and GPT2-M, Antonello and Huth (2023).

Further experiments challenge predicting coding theory: with a technique similar to RGP, word representations from different transformers are mapped to activations from sequence-to-sequence models, that is LMs that can translate between English and German (*ENG* $\rightarrow$ *DE translation transfer performance*); this metric is found to well correlate with brain mapping scores ($R = 0.780$). Given the findings from Caucheteux et al. (2023), one could argue that the learning objective of translating between the two languages underlies the learned concepts as in the brain: within this experimental setting, it would be absurd as none of the subjects in the study speak fluent German. Alternatively, Antonello and Huth (2023) suggest that linguistic models that learned general enough features to transfer to translation do map well to brain activity.

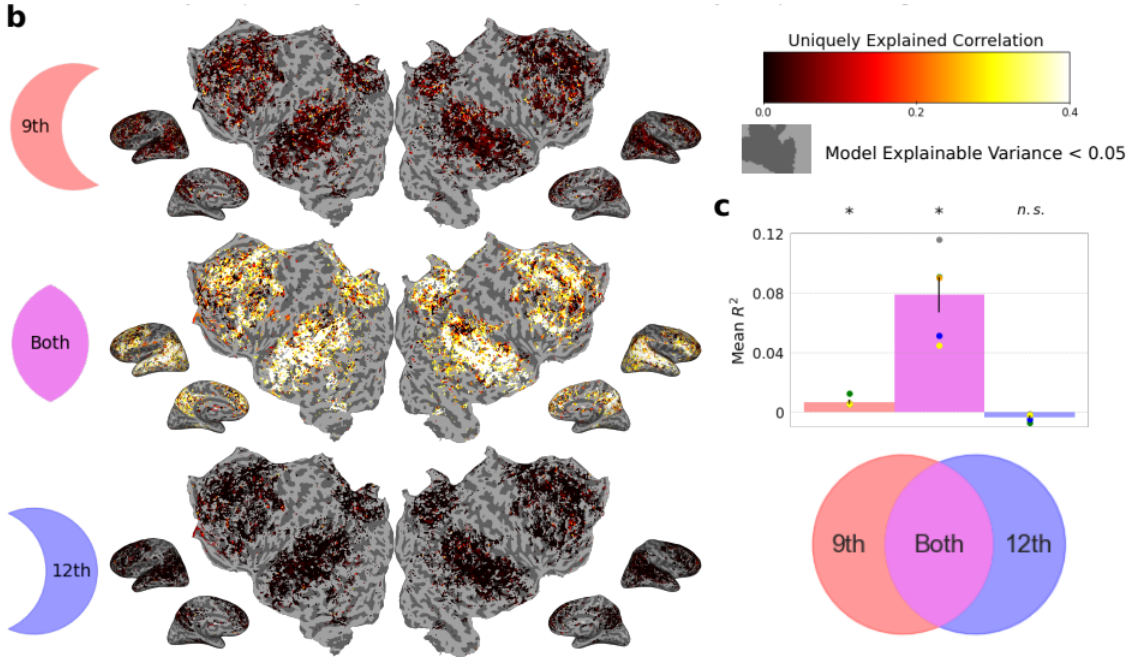

Figure 15. (B) voxel-wise contributions to brain response variance for each model (9th, 12th layer and both), (C) mean regression scores to brain response variance for the three models, Antonello and Huth (2023).

Finally, an analysis on the performance of different layers in GPT-2 models further tests the relationship between brain mapping and next-word prediction. As found by Caucheteux and King (2022), intermediate layers best predict brain activity, while the last ones best perform in language modeling, Figure 14 (left); this reflects the similarity between the inputs and activations in the earliest layers, as well as respectively the outputs with activations in the final layers. Degradation in brain mapping performance of the last layers, Figure 14 (right), should not occur if next-word prediction were the direct cause for similarity with neural activity. Additionally, with Variance Partition Analysis, activations in the 12th layer (last) are shown not to uniquely explain variance in brain responses in any cortical region, Figure 15.B, conversely to embeddings extracted from the 9th layer, which also report a higher regression score, Figure 15.C. In conclusion, we can argue that activations in the middle layers of GPT-2 may encode linguistic features that are agnostic to the training task and they're general enough to map to brain activity. Antonello and Huth (2023) suggests that representational generality may underlie this success (compatibly to Santoro et al. (2022) and Binder et

al. (2016)) and thus the next word prediction task just allows these concepts to emerge.

The reported observations are compatible with the idea that learned concept that well transfer across tasks and modalities are shared between computational models and the brain. Two questions follow: what learned features constitute such general concepts? How do sentence-level predictive tasks alter word representations and consequently brain mappings?

Reasoning properties and limitations of transformer models

An analysis on generalization and compositionality of computational features

Under the theory that brain processing is based on hierarchies, it is valid to assume that algebraic operations underlie the relationships between concepts: simple notions can be composed into more complex ones, e.g. opening a door implies walking towards it, pulling the handle and pushing it. This originates the idea of linguistic compositionality, which is at the base of natural language understanding research: in earlier attempts from Chomsky (1956), discrete sets of symbols and production rules were used to reproduce language production; later on distribution-based methods gained traction as they can learn the distribution of text from large corpora without explicit design. Arguably, this recent technique can learn to produce language fluently, that is to be linguistically productive, without necessarily constructing and conveying a meaning.

Early experiments on compositionality in neural networks have been conducted by Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018), which tested a recurrent neural network pre-trained for language modeling to classify the next word among two options: a grammatically correct one or a random inflection. The network achieved a largely above-chance-classification accuracy on long sequences, suggesting it may be able to learn linguistic rules over mere fluency from context. Lakretz et al. (2019) expanded on this work and found that predictions rely on the activation of few units and, along with another subnetwork, these are sensible to inputs with complex constituency trees. Lake and Baroni (2018) introduces SCAN: a benchmark to test compositionality in sequence2sequence models by translating commands into outputs,

e.g. "jump twice" = "jump jump". RNN performance is found to depend on the training data: with random splits, the network achieves near perfect accuracy but it's assumed to depend on fuzzy learned patterns. Splits omitting one primitive, e.g. the word "right", test the capability of the network to extrapolate to unseen commands e.g. by seeing examples of "left" opposite to "right" and "jump" associated with "left", it should understand and execute "jump right"; the performance of RNNs in this setting is below chance-level, supporting the idea that the learned patterns are brittle.

Dessì and Baroni (2019) introduce a novel architecture, based on convolutional layers and attention, drastically improving accuracy by circa 40%: the choice of the inductive biases, encoded in the architecture, determines better compositional capabilities. Subsequent experiments, however, reveal that prediction error is uniform across unseen patterns, raising doubts on the consistency of the learned patterns. Productivity without compositionality in neural networks is hypothesized also by Andreas (2019) in emergent communication: one network has to communicate to the other the object to classify through a discrete set of symbols, this task is shown to be solved by learning representations that do not compose with each other.

Transformers encode different inductive biases with respect to convolutional and recurrent neural networks: portions of input data are compared with a "key matching", or dictionary, paradigm. Ontanon, Ainslie, Fisher, and Cvicek (2022) tests various additions to the transformer architecture that can improve compositionality capabilities: relative position encoding, that is encoding each pair of input tokens with their distance, helps in tasks relying on position information e.g. maths; a copy decoder, which adds a cross-attention layer attending the transformer encoder's output, significantly improves performance in all compositional tasks; increased model size is shown not to generally improve performance, few exceptions involve tasks where prior knowledge is crucial on top of logical rules; weight sharing largely improves performance in tasks that require learning primitives e.g. "right"/"jump"/"run", using the same weights across depth increases the robustness of learned concepts across levels of abstraction; finally, simplifying the target outputs with intermediate sequence

representations (or formats) overall slightly improves performance on SCAN.

Arguably, the success of current language models is partially attributed to their linguistic fluency, that is their productivity. However, their capabilities in constructing concepts from simple notions and following logical rules are uncertain and currently at the center of NLP research. The choice of inductive biases, that is the logic by which information flows in the neural network, is arguably the most relevant factor.

Quantifying generality in learned language features

In the previous chapters I reviewed a set of features that gradually improved computational word embeddings to predict brain activity: activations from the middle layers of a transformer model seem to best abstract from both raw inputs and the training task; contextualized word representations best predict semantic domains linked to language as the main modality; high performance in language modelling and representation generality across different models are correlated with high brain mapping scores. One could thus suppose that transformer models generalize concepts beyond the training task and context: how can this property be measured? How does this affect similarity with brain's processing?

As previously argued, a language model may not be capable of learning the true meaning of a concept only from textual instances. However, Patel and Pavlick (2022) show that only with language, even "shallow" relations between concepts can hold in grounded domains i.e. representational spaces enriched with concrete world examples, this spatial property is defined as **isomorphism**. To test this hypothesis, GPT-2 and GPT-3 models of different sizes are provided with few examples of grounded concepts, e.g. the word "left" to explain the position of a tile in a text grid, instead of further training the model, the examples are included in the textual input (**in-context learning**, Brown et al. (2020)); subsequently, the models are asked (1) to predict grounded concepts in an unseen world example (of different shape or size), (2) to predict a previously unseen grounded concept. Three conceptual domains are considered: spatial (e.g. left, right), cardinal (e.g. nord, south), colour.

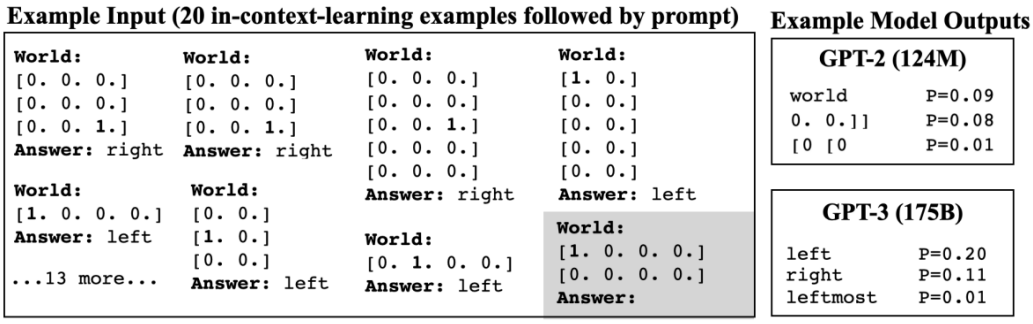

Figure 16. Model predictions for an unseen world instance of the concept "left", Patel and Pavlick (2022).

As GPT models are trained on large text corpora, it is possible that the prompted worlds have been already seen at training time in analogous forms, e.g. a matrix localizing a semantic position. To verify whether transformers perform actual reasoning over memorisation and pattern matching, as challenged by Dziri et al. (2023), the model is tested on "rotated" versions of the world examples that respect the structure of concepts: this corresponds to shifting by fixed amounts the position of the tiles in the matrices for the spatial and cardinal worlds. Randomly rotated worlds are eventually tested to verify the effects when isomorphism is not preserved.

	Top-1 Accuracy						**Top-3 Accuracy**					
	Spatial			**Cardinal**			**Spatial**			**Cardinal**		
	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.
R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.16	0.16	0.16	0.13	0.13	0.13
124 M	0.11	0.10	0.10	0.13	0.12	0.11	0.23	0.21	0.10	0.25	0.24	0.10
355 M	0.12	0.12	0.10	0.11	0.14	0.10	0.24	0.25	0.15	0.23	0.14	0.12
774 M	0.08	0.09	0.10	0.11	0.12	0.11	0.12	0.19	0.14	0.18	0.17	0.11
1.7 B	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.18	0.15	0.12	0.12	0.13
175 B	**0.45**	**0.44**	**0.16**	**0.43**	**0.46**	**0.18**	**0.76**	**0.75**	**0.19**	**0.88**	**0.76**	**0.21**

Table 7. Accuracy scores across domains and models in generalising to unseen worlds, Patel and Pavlick (2022).

To measure generalisation to unseen worlds, the models are asked to predict concepts for unseen scenes with 20 examples; top-1 and top-3 accuracy correspond to the fraction of cases when the model includes the correct guess in 1 or 3 generated

sentences, respectively. As baseline, the accuracy of random guesses are considered: **R-IV** for words randomly sampled from the model's vocabulary; **R-ID** for words randomly sampled from the in-domain words (e.g. north, south, in the spatial domain). A significant gap in performance is present between GPT-3 (175B parameters) and the rest of the models, including GPT-2 1.7B, GPT-2 774M and random guesses, Table 7; this is compatible with the common idea that GPT-3 well learns from context and its learned representations well transfer to new tasks, a small difference in performance between the original and the rotated worlds suggests that the model does not indeed memorize world layouts but rather semantic patterns.

To test performance with unseen grounded concepts, the models are prompted to predict novel concepts given some examples. For instance: in the spatial domain, to predict "right", grounded examples include world matrices for positions such as up, left, down-left; in the colour domain, different RGB triplets are presented along with labels from 6 primary colours or 57 colours from combinations, a novel color combination is then requested given the RGB code. Surprisingly, the accuracy achieved by the largest model, GPT-3, exceeds by a significant margin the scores from random guesses and smaller models in all the domains and transformation (original, rotated, random), Table 8. Similarly to the previous experiment, performance is preserved in predicting concepts from "rotated" world instances, which further suggests that logical relationships between e.g. colors, spatial references are transferred to isomorphic grounded spaces.

		Spatial			**Cardinal**			**Colours**		
		Original	Rotated	Random	Original	Rotated	Random	Original	Rotated	Random
Top-1 Accuracy	**R-IV**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	**0.16**	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.10	0.11	0.04	0.11	0.10	0.05	0.08	0.09	0.03
	355 M	0.10	0.10	0.04	0.10	0.11	0.06	0.06	0.07	0.04
	774 M	0.09	0.11	0.03	0.13	0.12	0.08	0.11	0.09	0.01
	1.5 B	0.14	0.14	0.12	0.13	0.14	0.10	0.10	0.09	0.06
	175 B	**0.28**	**0.27**	0.13	**0.30**	**0.29**	**0.08**	**0.23**	**0.21**	**0.11**
Top-3 Accuracy	**R-IV**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.13	0.12	0.07	0.09	0.09	0.08	0.06	0.05	0.04
	355 M	0.24	0.17	0.15	0.19	0.17	0.10	0.14	0.11	0.12
	774 M	0.19	0.24	0.12	0.17	0.15	0.11	0.15	0.16	0.14
	1.5 B	0.32	0.29	0.20	0.21	0.20	0.14	0.19	0.18	0.16
	175 B	**0.64**	**0.65**	**0.21**	**0.60**	**0.61**	**0.09**	**0.34**	**0.36**	**0.13**

Table 8. Prediction scores for unseen concepts in each domain and permutation, Patel

and Pavlick (2022).

It is worth mentioning that the prompted tasks are differently understood by models of varying sizes. When these fail to ground concepts, it is possible to distinguish conceptual errors by simple hallucinations: GPT-3 (175B) is shown to almost always generate "in-topic" answers, that is words belonging to the prompted domain (in 98% of the prompts), contrarily to smaller models (in 53% of the prompts). Moreover, a **grounding distance** is computed to quantify the closeness of transferred concepts to ground truth: e.g. in color domain, the distance between the predicted color "red" and ground truth "orange" is the euclidean distance between their respective RGB coordinates. Language models of increasing size tend to close this gap, Table 9.

	124M	**355M**	**774M**	**1.5B**	**175B**
C	328.3	309.5	209.6	190.7	96.3
R-IV	334.9	334.9	334.9	334.9	334.9
R-ID	174.9	174.9	174.9	174.9	174.9

True G	**Predicted** G **& Distance**
dark red	wine (76.5), light crimson (208.1) dark slate gray (144.7)
light green	beige (126.6), light sea green (129.7) cerulean (185.7), violet (262.6)

Table 9. (left) grounded distance of random guesses and GPT models (C) from the ground truth; (right) model predictions and relative grounding distances wrt. the truth, Patel and Pavlick (2022).

This study provides evidence that language models learn abstract representations that can be instanced with further information from the concrete world. As claimed by Patel and Pavlick (2022), a new approach to grounding language models could be proposed: instead of training a transformer from scratch with cross-modal data, it is sufficient that large models learn sufficiently rich conceptual structure to transfer to concrete instances more efficiently.

Challenging the intrinsic problem solving ability of transformer models

Arguably, transformers can learn concepts at different levels of abstraction, Caucheteux et al. (2023), with logical relationships, even shallow, that can be enriched with grounding, Patel and Pavlick (2022); Baroni (2019) provided evidence that in reasoning tasks, simple concepts (*primitives*) can be learned and executed from

instructions that combine them, but to understand the underlying computational rules is a task where neural networks perform inconsistently. This poses a major question: can neural networks learn to reason by construction? Put in the context of large language models, do transformers learn to perform algebraic operations or do they learn by rote?

Dziri et al. (2023) advance two theories: transformers achieve high performance only in solving low complexity operations, especially when (partial) examples are provided in the training set, suggesting that reasoning is approximated with pattern matching; as computation in transformers flows in one way, errors originated at intermediate steps of computation propagate and lead to incorrect outputs.

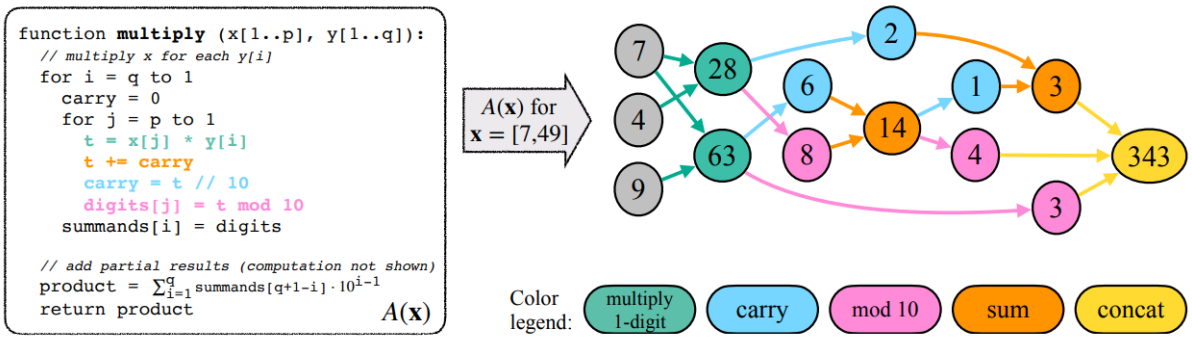

Figure 17. Source code (left) and computational graph (right) for the one-digit multiplication algorithm, Dziri et al. (2023).

These hypothesis are tested by considering deterministic algorithms, e.g. one-digit multiplication, represented by computational graphs: nodes are (partial) results of computation, edges are mathematical operations of different types; the depth of a computational graph and the average nodes per layer are proxies for complexity of the algorithm, Figure 17. Large language models of different sizes (GPT-3, Chat-GPT, GPT-4) are fine-tuned on (*question, answer*) pairs and evaluated with train-test validation on solving n-digits multiplication and Einstein's puzzle (assigning a subset of descriptions to each house in a list, given some constraints). For the first problem, the dataset is constructed by enumerating all the pairs of multiplying factors up to 4 digits, the models are evaluated on a subset not seen during training or examples with more

digits. Alternatively, the models are also trained with (*question, scratchpad*) pairs: a scratchpad is a step-by-step solution of a problem in written text. All the LLMs achieve near-perfect performance on the training split but fail to generalize at test time; scratchpads slightly improve performance, possibly prompted partial solutions may elicit the correct answer, Figure 18; success rate drastically decreases with more complex problem instances e.g. more digits, more elements to guess.

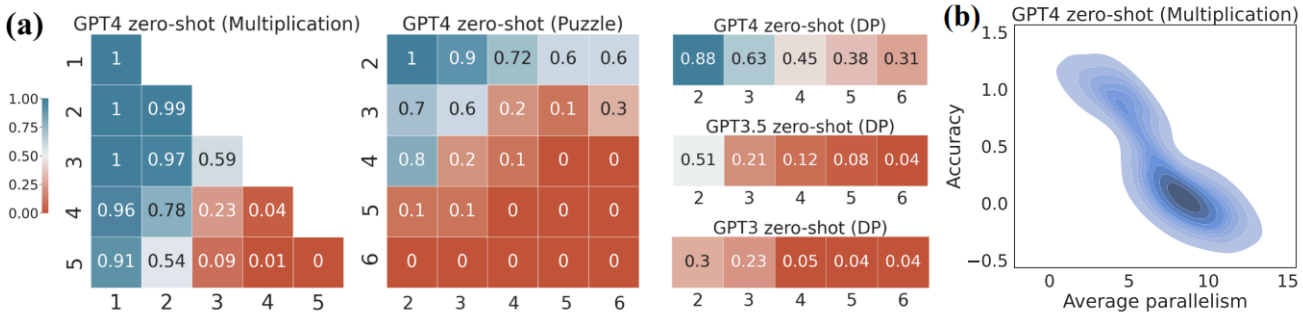

Figure 18. Zero-shot LM performance in two tasks, average parallelism is the ratio between the number of nodes and the reasoning depth of the graph, Dziri et al. (2023).

Despite the results suggest a lack of generalization of logic, transformers can to learn patterns between inputs and outputs: this is quantified by computing the Relative Information Gain (RIG) of a random variable given a conditioning stochastic process; in the multiplication task, high RIG is found between the first output digits and the first ones in the input factors, suggesting that numerical patterns could be learned. This hypothesis is tested by counting the occurrencies of each computational subgraph (each node with all its ancestors) in the training set: this quantity is high for all the correctly predicted (partial) results. Finally, the computational graph is analysed to localize the cause for incorrect predictions: local errors, when a node is incorrect but its parents are not; propagation errors, when some parent nodes are incorrect; restoration errors occur when all incorrect nodes precede a correct partial output. Propagation errors systematically outnumber local ones and they increase with deeper graphs: this means that not only the success rate of prediction decreases with problem complexity, but also that the model can learn single-step operations that fail to compose, leading to

propagation errors.

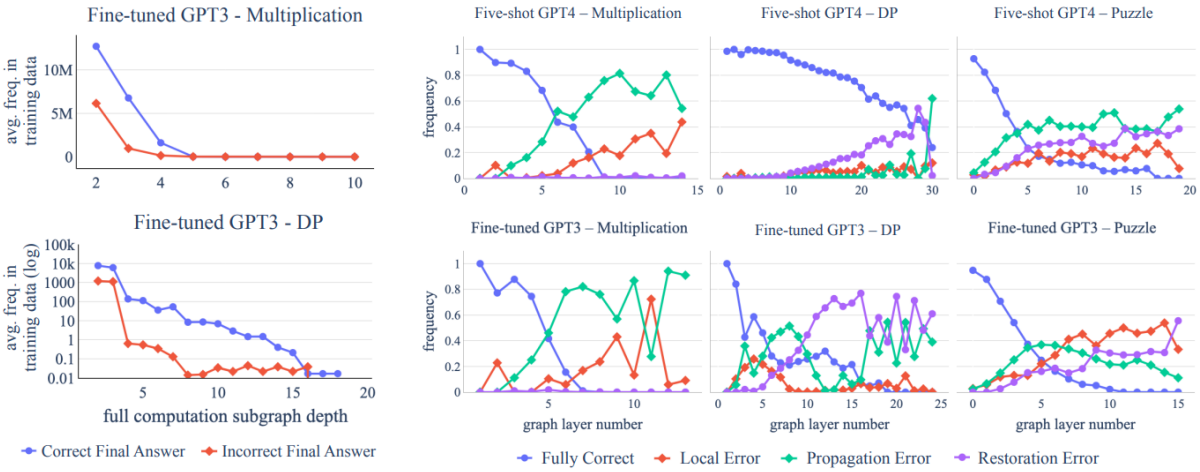

Figure 19. (left) average frequency of occurrency of test sub graphs in the training set wrt. subgraph depth; (right) frequency of nodes per depth level in each error category, Dziri et al. (2023).

The incapability of transformers in generalising algorithms can be traced to the training objective: next-word prediction allows memorisation over reasoning, that is shortcut learning; I conjecture that the lack of compositional skills may determine the divergence in information processing between large language models and the brain.

Conclusion

Arguably, transformer models constituted a turning point in deep learning: the attention mechanism is universal and applicable to any sequential task at scale. Current approaches on large-scale network pre-training take from computational linguistics: semantics are assumed to emerge from statistical properties in the inputs; this is linkable to learning in humans, which construct internal representations of the world through experience and observation.

Understanding concepts goes beyond communicating them fluently: to convey meaning, the speakers have to agree on a shared "world model" whose entities refer to concrete examples or not; reasoning implies understanding the relationships between these elements and applying logical operations, supporting the idea that information is hierarchical and composable. Research in the intersection between neuroscience,

linguistics, and artificial intelligence suggests that the brain operates by alternating observation with prediction of multiple time horizons.

The success of large language models is partially justified: beyond linguistic fluency, they can understand relationships between concrete and some abstract concepts, but fail at constructing new meanings from them; new evaluation techniques are necessary to assess their problem-solving capabilities over memorization. The similarity of computational representations to brain activity may be a proxy for the quality of learned concepts: models that predict multiple words in the future and demonstrate semantic comprehension better resemble human information processing.

These observations converge to one conclusion: it is necessary to shift to a cognitive approach in model training and evaluation, driving model design towards better task solvers.

References

Andreas, J. (2019). *Measuring compositionality in representation learning.*

Antonello, R., & Huth, A. (2023, 02). Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language*, 1-16. Retrieved from https://doi.org/10.1162/nol_a_00087 doi: 10.1162/nol$_{a0}$0087

Baroni, M. (2019). Linguistic generalization and compositionality in modern artificial neural networks. *CoRR*, *abs/1904.00157*. Retrieved from http://arxiv.org/abs/1904.00157

Binder, J., Conant, L., Humphries, C., Fernandino, L., Simons, S., Aguilar, M., & Desai, R. (2016, 06). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*. doi: 10.1080/02643294.2016.1147426

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from https://aclanthology.org/Q17-1010 doi: 10.1162/tacl$_{a0}$0051

Britton, B. K. (1978). Lexical ambiguity of words used in english text. *Behavior Research Methods & Instrumentation*, *10*(1), 1-7. Retrieved from https://doi.org/10.3758/BF03205079 doi: 10.3758/BF03205079

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). *Language models are few-shot learners.*

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). *Disentangling syntax and semantics in the brain with deep networks.*

Caucheteux, C., Gramfort, A., & King, J.-R. (2022, 09). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*. doi: 10.1038/s41598-022-20460-9

Caucheteux, C., Gramfort, A., & King, J.-R. (2023, 03). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*, 1-12. doi: 10.1038/s41562-022-01516-2

Caucheteux, C., & King, J.-R. (2022, Feb 16). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. Retrieved from https://doi.org/10.1038/s42003-022-03036-1 doi: 10.1038/s42003-022-03036-1

Chersoni, E., Santus, E., Huang, C.-R., & Lenci, A. (2021, November). Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, *47*(3), 663–698. Retrieved from https://aclanthology.org/2021.cl-3.20 doi: 10.1162/coli$_{a0}$0412

Chiu, B., & Baker, S. (2020, 12). Word embeddings for biomedical natural language processing: A survey. *Language and Linguistics Compass*, *14*. doi: 10.1111/lnc3.12402

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113-124. doi: 10.1109/TIT.1956.1056813

Davis, C., & Yee, E. (2018, 05). Features, labels, space, and time: factors supporting taxonomic relationships in the anterior temporal lobe and thematic relationships in the angular gyrus. *Language, Cognition and Neuroscience*, *34*, 1-11. doi: 10.1080/23273798.2018.1479530

Dessì, R., & Baroni, M. (2019, July). CNNs found to jump around more skillfully than RNNs: Compositional generalization in seq2seq convolutional networks. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3919–3923). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1381 doi: 10.18653/v1/P19-1381

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, *abs/2010.11929*. Retrieved from

https://arxiv.org/abs/2010.11929

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., ... Choi, Y. (2023). *Faith and fate: Limits of transformers on compositionality.*

Ettinger, A., Elgohary, A., & Resnik, P. (2016, August). Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 134–139). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W16-2524 doi: 10.18653/v1/W16-2524

Faass, G., & Eckart, K. (2013). Sdewac – a corpus of parsable sentences from the web. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language processing and knowledge in the web* (Vol. 8105, p. 61-68). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-40722-2_6

Friston, K. (1995). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, *2*, 56–78.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018, June). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1108 doi: 10.18653/v1/N18-1108

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., & Pietrini, P. (2001, 10). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, *293*, 2425-30. doi: 10.1126/science.1063736

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2201968119 doi: 10.1073/pnas.2201968119

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hollenstein, N., de la Torre, A., Langer, N., & Zhang, C. (2019, November). CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd conference on computational natural language learning (conll)* (pp. 538–549). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K19-1050 doi: 10.18653/v1/K19-1050

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*(16), 9379-9384. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.96.16.9379 doi: 10.1073/pnas.96.16.9379

Kaiser, D., Jacobs, A. M., & Cichy, R. M. (2022, 02). Modelling brain representations of abstract concepts. *PLOS Computational Biology*, *18*(2), 1-15. Retrieved from https://doi.org/10.1371/journal.pcbi.1009837 doi: 10.1371/journal.pcbi.1009837

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *CoRR*, *abs/2001.08361*. Retrieved from https://arxiv.org/abs/2001.08361

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th international conference on neural information processing systems - volume 1* (p. 1097–1105). Red Hook, NY, USA: Curran Associates Inc.

Lake, B. M., & Baroni, M. (2018). *Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks.*

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019, June). The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies,*

volume 1 (long and short papers) (pp. 11–20). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1002 doi: 10.18653/v1/N19-1002

Lecun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech and time series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 255–258). The MIT Press.

Lenci, A. (2018, 02). Distributional models of word meaning. *Annual Review of Linguistics*, *4*. doi: 10.1146/annurev-linguistics-030514-125254

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th international conference on neural information processing systems - volume 2* (p. 2177–2185). Cambridge, MA, USA: MIT Press.

Liao, J., Chen, X., & Du, L. (2023). *Concept understanding in large language models: An empirical study.* Retrieved from https://openreview.net/forum?id=losgEaOWIL7

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.*

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. Retrieved from https://www.science.org/doi/abs/10.1126/science.1152876 doi: 10.1126/science.1152876

Ontanon, S., Ainslie, J., Fisher, Z., & Cvicek, V. (2022, May). Making transformers solve compositional tasks. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 3591–3607). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.251 doi: 10.18653/v1/2022.acl-long.251

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). Understanding the exploding gradient

problem. *CoRR*, *abs/1211.5063*. Retrieved from http://arxiv.org/abs/1211.5063

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=gJcEM8sxHK

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1162 doi: 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1202 doi: 10.18653/v1/N18-1202

Price, A., Bonner, M., Peelle, J., & Grossman, M. (2015, 02). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *35*, 3276-84. doi: 10.1523/JNEUROSCI.3446-14.2015

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Santoro, A., Lampinen, A., Mathewson, K., Lillicrap, T., & Raposo, D. (2022). *Symbolic behaviour in artificial intelligence.*

Stephen, F., & Ramazan, G. (2010). An introduction to textual econometrics. In (p. 139). Retrieved from https://api.semanticscholar.org/CorpusID:168495426

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., & MacDiarmid, M. (2023). *Activation addition: Steering language models without optimization.*

Utsumi, A. (2020, 06). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*. doi: 10.1111/cogs.12844

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Vigliocco, D. P. V. G., Gabriella, & Gareth, M. (2007). *The Oxford Handbook of Psycholinguistics*. Oxford University Press. Retrieved from https://doi.org/10.1093/oxfordhb/9780198568971.001.0001 doi: 10.1093/oxfordhb/9780198568971.001.0001

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020, November). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 7222–7240). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.586 doi: 10.18653/v1/2020.emnlp-main.586

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*(51), 20754-20759. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1117807108 doi: 10.1073/pnas.1117807108

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014, October). Aligning context-based statistical models of language with brain activity during reading. In

Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 233–243). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1030 doi: 10.3115/v1/D14-1030

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). *Emergent abilities of large language models.*

Wikipedia. (2023). *Zipf's law.* Retrieved from

Yee, E., & Thompson-Schill, S. (2016, 06). Putting concepts into context. *Psychonomic Bulletin Review*, *23*, 1015-1027. doi: 10.3758/s13423-015-0948-7