COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13388

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13388

Modeling Brain Representations of Words' Concreteness in Context Using GPT-2 and Human Ratings

Andrea Bruera,[a,b] Yuan Tao,[c] Andrew Anderson,[d] Derya Çokal,[e] Janosch Haber,[a,f] Massimo Poesio[a,g]

[a]*School of Electronic Engineering and Computer Science, Cognitive Science Research Group, Queen Mary University of London*
[b]*Lise Meitner Research Group Cognition and Plasticity, Max Planck Institute for Human Cognitive and Brain Sciences*
[c]*Department of Cognitive Science, Johns Hopkins University*
[d]*Department of Neurology, Medical College of Wisconsin*
[e]*Department of German Language and Literature I-Linguistics, University of Cologne*
[f]*Chattermill, London*
[g]*Department of Information and Computing Sciences, University of Utrecht*

Received 19 April 2023; received in revised form 12 September 2023; accepted 27 October 2023

Abstract

The meaning of most words in language depends on their context. Understanding how the human brain extracts contextualized meaning, and identifying where in the brain this takes place, remain important scientific challenges. But technological and computational advances in neuroscience and artificial intelligence now provide unprecedented opportunities to study the human brain in action as language is read and understood. Recent contextualized language models seem to be able to capture homonymic meaning variation ("bat", in a baseball vs. a vampire context), as well as more nuanced differences of meaning—for example, polysemous words such as "book", which can be interpreted in distinct but related senses ("explain a book", information, vs. "open a book", object) whose differences are fine-grained. We study these subtle differences in lexical meaning along the concrete/abstract dimension, as they are triggered by verb-noun semantic composition. We analyze functional magnetic resonance imaging (fMRI) activations elicited by Italian verb phrases containing nouns whose interpretation is affected by the verb to different degrees. By using a contextualized language model and human concreteness ratings, we shed light on where in the brain such fine-grained meaning

Correspondence should be sent to Andrea Bruera, Lise Meitner Research Group Cognition and Plasticity, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig 04103, Germany. E-mail: bruera@cbs.mpg.de

variation takes place and how it is coded. Our results show that phrase concreteness judgments and the contextualized model can predict BOLD activation associated with semantic composition within the language network. Importantly, representations derived from a complex, nonlinear composition process consistently outperform simpler composition approaches. This is compatible with a holistic view of semantic composition in the brain, where semantic representations are modified by the process of composition itself. When looking at individual brain areas, we find that encoding performance is statistically significant, although with differing patterns of results, suggesting differential involvement, in the posterior superior temporal sulcus, inferior frontal gyrus and anterior temporal lobe, and in motor areas previously associated with processing of concreteness/abstractness.

Keywords: Semantics; fMRI; Computational linguistics; Polysemy; Concreteness; Semantic composition; Machine learning; Language models

1. Introduction

1.1. Meaning variation and semantic composition

Most words are **ambiguous**—they can be interpreted differently depending on the context. However, not all words are ambiguous in the same way. Much of the research on lexical semantic interpretation in linguistics, neuroscience, psychology, and natural language processing (NLP) has focused on **homonyms**—words like *bat* that have interpretations that are completely unrelated (a flying mammal vs. an implement used in games such as baseball: Lyons, 1977; Pinkal, 1995; Rodd, Gaskell, & Marslen-Wilson, 2002). For other words, however, the differences between the different interpretations are much more fine-grained. A notorious example of fine-grained variation are cases of semantic **polysemy** (Apresjan, 1974; Cruse, 1992; Falkum & Vicente, 2015; Haber & Poesio, 2023; Lyons, 1977; Nerlich, Todd, Herman, & Clarke, 2003; Pinkal, 1995; Pustejovsky, 1998). The word *book* is a classic example of a polysemous word: its interpretation can vary depending on the context, but these different interpretations are intuitively related. As an example, in *open the book*, the noun *book* is used to refer to book as a physical object, whereas in *summarize the book*, the same noun is used to refer to an abstract object, the content of the book. Theoretical linguistics has extensively investigated how the process of **semantic composition** triggers variation in lexical meaning, where the interpretation of words, like the polyseme *book*, is affected by its (syntactic) context (Frege, 1892; Montague, 1973; Pustejovsky, 1998; Partee, 2008; Pustejovsky, 2011). The theories of composition proposed in this line of research explain the interplay between syntactic and semantic interpretation—that is, how linguistic units combine systematically in order to convey complex meaning. In the study reported in this paper, we investigate the effect of semantic composition on fine-grained meaning variation, including but not limited to polysemy, using brain evidence.

1.2. Semantic composition in neuroscience

We focused on compositionality effects on the lexical representation in Italian verb-noun phrases resulting in the contrast between *open the book* versus *summarize the book* discussed

earlier (*aprire il libro* vs. *riassumere il libro*; in the following, we will report our example phrases in English, instead of their original Italian version, to facilitate understanding). In such constructions, the level of concreteness of the noun is determined by the verb: in the case of *open the book* / *summarize the book*, for instance, the semantic type of the noun is refined to either physical object or information. These phrases are particularly interesting because they allow us to investigate how the brain handles fine-grained meaning variation and how this interacts with semantic composition: in other words, if, and how, the process of semantic composition alters the representation of its underlying parts (the verb and the noun), and where these representations are located in the brain.

The meaning variations triggered by semantic composition in verb-noun phrases have been extensively investigated in the linguistic literature, but not in neuroscience, where adjective-noun phrases have been studied much more frequently (Bemis & Pylkkänen, 2011; Fritz & Baggio, 2020; Kochari, Lewis, Schoffelen, & Schriefers, 2021; Murphy et al., 2022; Pylkkänen, 2020), and especially not using brain encoding methods, which we argue could offer insights into their neural representation. In the few cases where compositionality involving verbs and nouns has been studied in the cognitive neuroscience of language, this has been mostly looked at as a syntax or syntax-semantics interface phenomenon. In such experiments, the authors modulated the syntactic structure (Zaccarella, Meyer, Makuuchi, & Friederici, 2017), changed the semantic roles for the nouns (Frankland & Greene, 2020), and/or looked at the effects of positive/negative polarity (Zhang & Pylkkänen, 2018) or argument saturation or modification (Westerlund, Kastner, Al Kaabi, & Pylkkänen, 2015). Little attention has been paid to keeping syntactic or syntax-semantic interface elements unchanged, and instead studying verb-noun composition by modulating meanings (with the only exceptions, to our knowledge, of (Husband, Kelly, & Zhu, 2011; Sakreida et al., 2013)). The semantic modulation approach was so far only adopted for adjective-noun composition cases (Fyshe, Sudre, Wehbe, Rafidi, & Mitchell, 2019; Honari-Jahromi, Chouinard, Blanco-Elorrieta, Pylkkänen, & Fyshe, 2021; Pylkkänen, 2020).

1.3. Brain encoding

Another novelty in our work is that we use a multivariate brain encoding approach, in combination with computational models. Previous studies of the interplay between semantic composition and polysemic variation in the interpretation of lexical items in cognitive neuroscience (Klepousniotou, Pike, Steinhauer, & Gracco, 2012; Klepousniotou, Gracco, & Pike, 2014; Lukic, Meltzer-Asscher, Higgins, Parrish, & Thompson, 2019; MacGregor, Bouwsema, & Klepousniotou, 2015; Mollica et al., 2020; Pylkkänen, Llinás, & Murphy, 2006; Pylkkänen & McElree, 2007; Pylkkänen, 2020) have all used univariate methods (e.g., looking at differences in BOLD activation across conditions), whereas multivariate analyses afford higher sensitivity and, more importantly, a way to test competing accounts (Hebart & Baker, 2018; Naselaris & Kay, 2015).

Brain encoding studies, where machine learning is used to predict patterns of brain activity by learning functions from computational representations—for example Abraham et al. (2014), Grootswagers, Wardle, & Carlson (2017), Haxby et al. (2001), Haxby, Connolly,

Guntupalli, & others (2014), Haynes (2015), Kragel, Koban, Barrett, & Wager (2018), Lemm, Blankertz, Dickhaus, & Müller (2011), Naselaris, Kay, Nishimoto, & Gallant (2011), Pereira, Mitchell, & Botvinick (2009), Rybář & Daly (2022)—have recently started to make use of vectorial models of meaning proposed in NLP that have been shown to capture an extremely wide range of information involved with semantic processing (for comprehensive reviews, see Hale et al., 2022; Murphy, Wehbe, & Fyshe, 2018). In encoding, vectorial semantic representations open new possibilities for the investigation of semantic processing in the brain (Bruffaerts et al., 2019; Diedrichsen & Kriegeskorte, 2017; Kay, 2018; Kriegeskorte, Mur, & Bandettini, 2008; Naselaris & Kay, 2015). By predicting the brain activation patterns for concepts, encoding makes it possible to directly compare competing models of cognitive phenomena by looking at their relative fit with brain processing (Naselaris et al., 2011). In this framework, vectorial semantic representations can offer novel insights with respect to the interpretation of the neural bases of semantic processing. Brain encoding methods involving vectorial semantic representations have been used to study various aspects of the interpretation of linguistic meaning: single-word or -concept meaning (Anderson, Murphy, & Poesio, 2014; Mitchell et al., 2008; Murphy, Baroni, & Poesio, 2009; Murphy et al., 2011; Pereira et al., 2018) or features (Kaiser, Jacobs, & Cichy, 2022; Sudre et al., 2012), whole-sentence meaning (Anderson et al., 2021; Jat, Tang, Talukdar, & Mitchell, 2019), teasing apart syntax and semantics (Caucheteux, Gramfort, & King, 2021), linguistic meaning in naturalistic contexts, such as narratives (Caucheteux & King, 2022; Dehghani et al., 2017; Goldstein et al., 2022; Wehbe et al., 2014) and movie transcripts (Vodrahalli et al., 2018), studies of noun-adjective composition (Fyshe et al., 2019; Honari-Jahromi et al., 2021), and the investigation of metaphors (Djokic, Maillard, Bulat, & Shutova, 2020). However, there have been no studies using computational language models from NLP to further our understanding of how the brain processes fine-grained meaning variation triggered by semantic composition. This is in large part because until recently, the prevalent vectorial models of lexical meaning—so-called **distributional semantics** representations discussed next (Baroni, Bernardi, & Zamparelli, 2014; Boleda, 2020; Clark, 2015; Camacho-Collados & Pilehvar, 2018; Erk, 2012; Griffiths, Steyvers, & Tenenbaum, 2007; Lund & Burgess, 1996; Landauer & Dumais, 1997; Lenci, 2018; Turney & Pantel, 2010)—assigned to phrases the same meaning in all contexts. However, this situation has radically changed recently. In this paper, we aim to fill this gap deploying more recent distributional models that do take multiplicity of meaning into account.

1.4. Language models and semantic composition

The objective of distributional semantics is to develop data-driven methods—these days typically called **language models**—for associating words to high-dimensional vectors that represent their meanings/usages using word co-occurrence information extracted from large collections of texts (Boleda, 2020; Harris, 1954; Turney & Pantel, 2010). Until recently, this line of research had focused on the development of what are usually called **static** language models (Apidianaki, 2022; Bojanowski, Grave, Joulin, & Mikolov, 2017; Griffiths et al., 2007; Lund & Burgess, 1996; Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014)—models which assign a single

interpretation to lexical items irrespective of context.[1] Such models would assign the same vectorial representation to *book* both in the phrase *open the book* and in the phrase *summarise the book*. Recently, however, the situation has changed with the development of **contextualized** language models, such as ELMO or BERT (Apidianaki, 2022; Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018; Radford et al., 2019). These are language models which assign to words representations that capture their linguistic meaning in a specific context—such models would assign *different* vectorial representations to *book* in the two contexts we are comparing. These models are suited to modeling meaning variation in context, opening, therefore, new opportunities for studying fine-grained composition phenomena within the brain encoding framework. But so far, no study has applied these models to investigating the effect of semantic composition on the interpretation of fine-grained variation in lexical meaning, like those happening in polysemic words. This is the primary objective of this work.

Language models have been used in cognitive research in different ways, depending on the characteristics of their representations (Günther, Rinaldi, & Marelli, 2019). Static language models, in which each word is associated with a single vector, have been interpreted as models of semantic memory (Lund & Burgess, 1996; Landauer & Dumais, 1997; Kumar, 2021). By contrast, contextualized language models, where what is represented are words appearing in linguistic contexts such as sentences, have been taken to be general models of semantic processing of both words and sentences (Caucheteux & King, 2022; Lenci, Sahlgren, Jeuniaux, Cuba Gyllensten, & Miliani, 2022). These models are specifically meant to capture semantic representations beyond specific semantic dimensions such as concreteness. Therefore, they can provide original insights with respect to brain processing of verb-noun semantic composition.

1.5. Summary of the analyses

We conducted two separate sets of encoding analyses of verb-noun semantic composition.[2]

For the first set of analyses, we considered the whole language network in the brain. By means of encoding, we looked at which models of lexical meaning and which models of semantic composition best capture brain processing of verb-noun semantic composition. Regarding lexical meaning, we compared a contextualized language model and "cognitive models" based on subjects ratings. For semantic composition, we contrasted two methods of composing representations, based, respectively, on single words and full phrases (described in detail in Section 4).

In the second set of encoding analyses, we focused on individual areas. In this case, we considered a number of regions of interest which have been previously argued to play a role in semantic composition in the brain (Pylkkänen, 2020; Sakreida et al., 2013). We investigated where in the brain we could most reliably use the concreteness of the phrases to predict brain signals, an indication that this type of information can be associated with activity in that area. Within the language network (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010), we considered the left anterior temporal lobe (**ATL**), the left posterior superior temporal sulcus (**pSTS**), and the left inferior frontal gyrus (**IFG**); outside of the language network, the bilateral ventro-medial pre-frontal cortex (**vmPFC**) and a bilateral

set of motor areas including the supplementary motor cortex (SMC) and the precentral gyrus (**motor-areas**; Sakreida et al., 2013).

1.6. Computational and cognitive models of lexical meaning and phrasal meaning

Our encoding analyses were carried out in relation to two types of models of meaning and, for each type of model, two ways of using such models to obtain a meaning for phrases.

The first type of model of meaning was a distributional model, namely, a contextualized language model, **ITGPT2**. **ITGPT2**, an adaptation for the Italian language of GPT2 (de Vries & Nissim, 2021; Radford et al., 2019), is a contextualized model which has been shown to excel at modeling semantics in the brain (Caucheteux & King, 2022; Schrimpf et al., 2021).

Such a model can be used in two different ways to compute the representation of phrases. Cognitive scientists contrast "simple" composition (also called "classical" composition in De Almeida et al., 2016, and "pure" composition in Fodor & Lepore, 2000), where the lexical representations of the constituents are *not* modified during composition to obtain a representation for the phrase, with "complex" composition (also called "enriched" composition in De Almeida et al., 2016). In this case, the meaning of the phrase is more than the meaning of its parts, and results from an operation on the semantic interpretation of the constituents which may involve a transformation of these interpretations (De Almeida et al., 2016; Goldberg, 1995; Pustejovsky, 1998). Contextualized language models are inherently models of complex composition, as the semantic representation of the lexical constituents is modified according to the linguistic context. For such models, the phrasal interpretation we use is directly computed by the model (we will call it **phrase ITGPT2**; details on the methodology used are given in Section 3). Contextualized language models can nevertheless be adapted to obtain generalized representations for individual lexical items, which are abstracted from specific linguistic contexts, like static models used to do (Apidianaki, 2022; Bommasani, Davis, & Cardie, 2020; Vulić, Ponti, Litschko, Glavaš, & Korhonen, 2020). Such single-word vectors can be then composed through a mechanism of simple composition, as they represent individual words in isolation. In this case, the semantic representation of a phrase is simply given by the average of the semantic representations of its parts (in the following, **single-words ITGPT2**; again, see Section 3).

The materials in this experiment were designed to study how semantic composition affects one specific semantic property: concreteness (see Section 3.1). Thus, the representations obtained from the contextualized language model were contrasted with what we called **cognitive models**, directly encoding information about concreteness provided by human subjects. Among these, the representation of phrases according to complex semantic composition (see above) is made of concreteness ratings provided by human subjects for the *full* phrases (**phrase concreteness**; see Section 3.1 for details on data collection). This models the assumption that raters carry out themselves the complex composition of the lexical items during the rating task (Naselaris et al., 2011), then representing the result through the concreteness rating given to each phrase. By contrast, in the simple composition model, a phrase's concreteness is the average of each individual word's concreteness (**single-words concreteness**)—a procedure that does not modify the semantic representations during composition.

1.7. Experimental design and stimulus selection

We modulate the semantics of the stimuli along a concreteness gradient, by varying the noun (polysemic or not) and the verb (requiring an object of a specific semantic type—abstract or concrete—or not).

We considered three such modes of composition, to achieve a systematic coverage of the constraints on interpretation imposed by the verb on the concreteness of the noun. In order to allow direct comparisons among each mode of composition, we used a balanced set of nouns belonging to one of the two semantic types involved in dot-objects such as *book* —namely, physical objects and information.

The first type of semantic composition we consider is **transparent** semantic composition, where a verb and a nonpolysemic noun have the same semantic type in terms of their concreteness (e.g., *open the envelope* for physical object and *explain the idea* for informational content), and therefore, the meaning of the phrase emerges transparently from the combination of the two parts (Bemis & Pylkkänen, 2011; Baggio, Van Lambalgen, & Hagoort, 2012; Jackendoff, 1997; Kamp & Partee, 1995; Partee, 2008; Pylkkänen & McElree, 2006; Pylkkänen, 2008). Previous work on the distributional properties of concrete and abstract nouns and verbs in large corpora indicates that this is the statistically dominant case (Frassinelli, Naumann, Utt, & m Walde, 2017; Frassinelli & Im Walde, 2019; Naumann, Frassinelli, & Schulte im Walde, 2018).

The second type of composition we considered is **sense selection** (also called "semantic type coercion") typical of polysemic contexts, whereby the verb selects the relevant sense of a polysemic noun—the so-called **dot-object** (see, e.g., Pustejovsky, 1998). Given a linguistic context, the interpretation of the noun can alternate between an abstract sense (informational content, with lower concreteness) and a concrete sense (physical object, with higher concreteness)—which results in different semantic types being assigned to the same noun when found in different phrases (Baggio, Choma, Van Lambalgen, & Hagoort, 2010; Frisson, 2015; Goldberg, 1995; Haber & Poesio, 2021; Jackendoff, 1997; Katsika, Braze, Deo, & Piñango, 2012; Kuperberg, Choi, Cohn, Paczynski, & Jackendoff, 2010; Lauwers & Willems, 2011; Pustejovsky, 1991, 1998; Pylkkänen, 2008; Pustejovsky et al., 2010; Pustejovsky, 2011; Pylkkänen, 2020; Zarcone, McRae, Lenci, & Padó, 2017).

Finally, we use phrases involving **light-verb** semantic composition, where the verb is a light verb, such as *have*. This type of verbs mostly plays a grammatical role, with very little to no semantic relevance in its context (Brugman, 2001; Briem et al., 2009; Butt, 2010; Wittenberg, Paczynski, Wiese, Jackendoff, & Kuperberg, 2014; Wittenberg & Levy, 2017), entailing that the interpretation of the phrase relies almost entirely on the noun and does not indicate a clear action (e.g., *have the envelope* vs. *have the idea*).

1.8. Summary of the results

Our results show that the two "complex" models of phrasal meaning (concreteness ratings on phrases and the full-phrase contextualized language model) capture better than "simple" models brain processing of semantic composition and can predict brain data with

statistical significance. We also find that, within the sense selection cases of semantic composition, phrasal models can discriminate among brain responses to different senses of some, but not all, nouns referring to dot-objects. This supports a view of semantic composition as a complex, holistic process operating on whole phrases and which strongly interacts with the semantics of the parts involved—in our case, the meaning of the verbs and the nouns. Also, we find that phrase ITGPT2 achieves better results than both simple composition models (single-words ITGPT2 and single-words concreteness) and close to those obtained with phrase concreteness judgments elicited from human subjects. This finding confirms that contextualized language models are able to capture to a surprising extent brain processing of semantics and semantic composition. However, overall the performance of the full-phrase ITGPT2 representations is slightly lower than that obtained using full-phrase human concreteness judgments. We interpret this as an indication that, while ITGPT2 can capture to a surprising extent brain processing of semantics and semantic composition, its ability to explain fine-grained meaning variation triggered by semantic composition as they happen in the brain leaves room, at least in its current form, for improvement.

Our results on individual areas indicate that, within the language network, the left pSTS and the left IFG are most consistently and strongly involved with all three types of semantic composition, also for polysemous nouns; however, we also find that all areas are involved to some degree in the processing of each mode of composition. We interpret this result as evidence that different linguistically defined cases of semantic composition elicit different patterns of brain activity, and therefore, involve brain areas distinctly depending on the cognitive resources required—a view which may help in reconciling apparently conflicting results from previous literature. Outside of the language network, instead, we find that motor areas contain information with respect to the representation of semantic composition. This confirms previous literature on neural processing of concreteness, which found activation of motor areas also with linguistic stimuli (Sakreida et al., 2013), and is compatible with an integration of linguistic and motor representations during semantic composition.

2. Semantic composition in the brain

A number of studies have investigated the areas of the brain in which semantic composition takes place. Earlier work focused on individual areas (**regions of interest —ROIs**), but in recent work, there has been a shift toward considering **networks**. We briefly summarize this work here.

2.1. Regions of interest

The brain regions associated with semantic composition in previous research are visualized in Figs. 1 (the language network) and 2 (for areas outside of the language network). All these regions were considered in the current study.

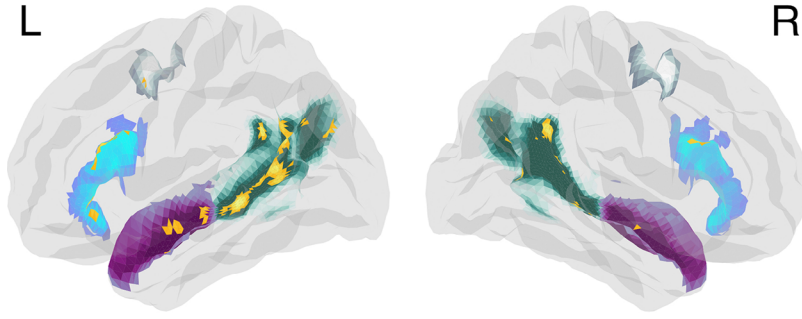

Fig. 1. Visualization of the regions of interest and selected voxels of the language network. For the subparts of the language network used for the ROI analyses (reported in Sections 6.3 and 6.4), colors correspond to each area's bar in Figs. 8 and 9 (ATL in purple, pSTS in green, and IFG in cyan). We report in yellow the top 25% most stable features (across all subjects) obtained for the encoding analysis from the language network using the stability selection procedure described in Section 3.2.7. They show a clear left-lateralization, with most voxels found in the pSTS and some in the ATL and the IFG as well. Brain areas are projected on the *fsaverage* cortical surface provided by FreeSurfer (Fischl, 2012).

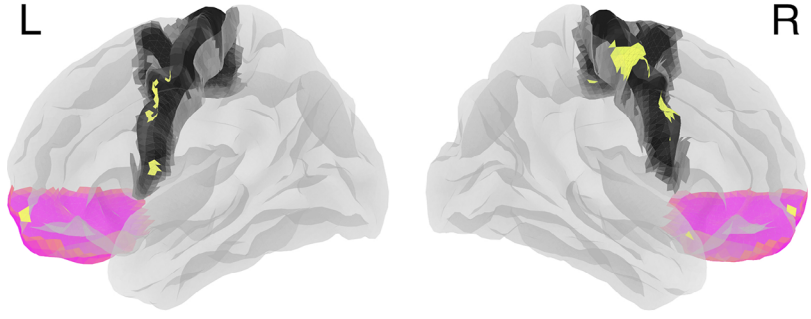

Fig. 2. Visualization of the regions of interest and selected voxels outside of the language network. Since these areas were used for the ROI analyses (reported in Sections 6.3 and 6.4), colors correspond to each area's bar in Figs. 8 and 9 (vMPFC in pink and motor areas in black). We report in yellow, for each area, the top 25% most stable features obtained using the stability selection procedure described in Section 3.2.7. For motor areas, the most stable features show some degree of right lateralization, whereas, within the vMPFC, the most stable features are located within the frontal pole. Brain areas are projected on the *fsaverage* cortical surface provided by FreeSurfer (Fischl, 2012).

2.2. *Left inferior frontal gyrus*

The left IFG has been proposed in Husband et al. (2011), the only previous work investigating dot-objects and sense selection with univariate fMRI analyses, as the main brain structure supporting sense selection processes triggered by semantic composition. Similarly, in Sakreida et al. (2013), the only other study looking at effects of verb-noun semantic composition in the brain, the left IFG was found to selectively respond to this modulation. Also, more generally, this brain area has been found to respond differently to abstract and concrete concepts (Binder, Westbury, McKiernan, Possing, & Medler, 2005; Bucur & Papagno,

2021; Della Rosa, Catricalà, Canini, Vigliocco, & Cappa, 2018), and to be associated with composition processes (Schell, Zaccarella, & Friederici, 2017).

2.3. Left anterior temporal lobe

The left ATL has been consistently shown to be a central hub for semantic processing of individual concepts (for a review, see Lambon Ralph, Jefferies, Patterson, & Rogers, 2017). In Pylkkänen (2020), which summarizes earlier Magnetoencephalography (MEG) studies, this area has also been proposed as the main locus for so-called conceptual combination in the brain.

2.4. Left posterior superior temporal sulcus

The left pSTS, which has been traditionally associated with a number of cognitive processes (Hein & Knight, 2008), including semantic (Price, Bonner, Peelle, & Grossman, 2015) and syntactic (Matchin & Hickok, 2020) combinatory processing, was recently shown in Murphy et al. (2022) to be strongly involved with semantic composition using intracranial electrocorticography recordings.

2.5. Ventro-medial pre-frontal cortex

We also considered the vMPFC, a brain area falling outside of the language network. The vMPFC has been implicated with processing of semantic composition in some MEG studies reviewed in Pylkkänen (2020), where its role was argued to emerge at a later time during semantic composition, in between language comprehension and production.

2.6. Motor areas

Finally, two motor areas, the precentral gyrus and the supplementary motor cortex (SMC), were found in Sakreida et al. (2013) to be strongly activated by verb-noun semantic composition with varying degrees of concreteness. The proposed explanation for such an activation, well outside of typical language areas, comes from the embodied cognition framework (Barsalouet al., 2008): in the strongest version, linguistic comprehension should trigger experiential simulations of the referents of the words in the brain (Fischer & Zwaan, 2008; Zwaan, 2004); in a softer version, which we adopt here, linguistic comprehension should at least involve to some extent modality-specific (in this case, motor) features which are integrated with supra-modal, linguistic information (Lambon Ralph et al., 2017), without postulating a full-on mental simulation of the action. If this were the case, then, phrases containing verbs and nouns of different concreteness should engage to different extents motor areas, depending on how much motor information is required for the linguistic comprehension of the verb-noun phrase.

2.7. Brain networks

Brain networks have emerged in recent years as a fundamental level of representation in cognitive neuroscience, going beyond individual brain areas (Suárez, Markello, Betzel,

& Misic, 2020). They allow to characterize complex cognitive processes, such as those related to language, at the same time taking into consideration individual variability within the network (Fedorenko et al., 2010; Friederici & Gierhan, 2013). A brain network can be defined either structurally—by looking at paths of physical connections among brain areas—or functionally—by reference to given cognitive processes, as a set of brain areas which have been found to respond collectively to them.

The **language network** is composed of brain areas individuated functionally in Fedorenko et al. (2010). Through a so-called "localizer task," the authors found brain areas selectively activated by meaningful sentences as opposed to lists of nonwords. More recently, it was shown that syntactic and semantic processes do not activate selectively specific areas, but rather engage the whole network (Fedorenko, Blank, Siegelman, & Mineroff, 2020).

In previous work, where pictures were used instead of words, authors conducted the encoding/decoding analyses using the whole brain instead of a brain network, or individual areas (e.g., Mitchell et al., 2008). This approach, however, has two main downsides. First, it tends to provide "salt-and-pepper" spatial distribution of the voxels retained after feature selection (cf. the original papers). Among these voxels, clusters falling within semantics- or language-specific networks may emerge, but a large amount of features are scattered all over the brain. This makes it hard to interpret results, as it is not clear what brain areas drive performance. Furthermore, as pointed out in Haynes (2015), it may even be that the patterns do not reflect information actually available to the brain during processing, because of anatomical constraints. Second, it has been shown in recent work looking at metaphorical meaning (Djokic et al., 2020) that, when stimuli are presented as words, a whole-brain analysis is not optimal when mapping between language models and the brain. To validate empirically our assumption, we also run a whole-brain analysis like the one used in Mitchell et al. (2008) and Pereira et al. (2018). The results are reported in Appendix A in the Supplementary Materials. Briefly, we find that accuracy is similar, but always slightly lower than that obtained with the language network (Appendix A, Figs. A.1 and A.2). Furthermore, the set of features retained after features selection is scattered all over the brain (Appendix A, Fig. A.3), empirically confirming our assumptions.

3. Methods

3.1. Stimuli

The experiment was carried out in Italian. The stimuli used in this study are Italian verb-noun phrases whose interpretation was judged by human subjects to occupy different positions on a concreteness gradient, depending on how a direct object (a noun referring either to a physical object, a piece of information, or a dot-object) is combined with a predicate (a verb).

Our 42 stimuli cover the three cases of semantic composition (**sense selection**, **transparent composition**, and **light-verb composition**); for each case, we selected 14 basic verb-noun (V+N) phrases, with seven phrases involving a noun belonging to the (abstract) semantic type **information**, and seven phrases involving a noun referring to a (concrete) **physical object**

(see stimulus selection procedure below). In the sense selection cases, the nouns (polysemous, "dot-object" cases) were the same across the two semantic types (e.g., *open/summarize a book*). These phrases are minimal examples of composition, involving only the minimum necessary amount of lexical units for composition to take place (Bemis & Pylkkänen, 2011; Fritz & Baggio, 2020; Kochari et al., 2021; Murphy et al., 2022; Pylkkänen, 2020), therefore, allowing to investigate from as close as possible its neural signature.

The 42 stimuli were obtained as follows. We came up with an initial pool of verb-noun phrases following the criteria discussed below. Those phrases were then rated by human subjects ($n = 36$) in terms of familiarity and concreteness. The final set of stimuli were selected so as to be matched in length, number of phonemes and familiarity for the object-information contrast within each composition case.

In the case of the sense selection stimuli, several nouns that can alternate between an abstract and a concrete meaning like *book*, and several verbs that can coerce the meaning of such nouns into either the abstract or the concrete interpretation were selected. Verb phrases combining these nouns and verbs were constructed, excluding those which had no meaning or were ambiguous. The final set of stimuli was built from four nouns (*book*, *magazine*, *catalogue*, *sketch*), and six coercing verbs: three referring to a physical action requiring as a direct object a physical object (*open*, *pick*, *give* (as a present)), three requiring as a direct object an information-related object (*explain*, *consult*, *present*).

In the transparent composition stimuli, the verb-noun phrases contained the same coercing verbs as the sense selection stimuli (i.e., verbs clearly requiring an information- or physical-object) and nouns unambiguously referring to information- or concrete-objects.

The nouns denoting physical objects are *parcel*, *ticket*, *flower*, *coin*, *ball*, *parcel*, *envelope*. The information-denoting nouns are *reason*, *word*, *problem*, *question*, *expert*, *program*.

Finally, in the light-verb stimuli, words which contained nouns of furniture and abstract information were selected and combined with the verbs (information: *opinion*, *judgment*, *idea*, *story*, *reason*; physical objects: *desk*, *table*, *sofa*, *wardrobe*, *chair*; verbs: *have*, *change*, *provide*). Common household furniture were chosen because they are frequently seen with book-type objects in real-world scenarios.

Phrases referring to physical objects and informational content were matched, within each mode of composition, for number of letters and phonemes across phrases involving physical objects and information (number of letters: $p_{Senseselection} = .65$; $p_{lightverb} = .44$; $p_{transparent} = .2$. number of phonemes: $p_{Senseselection} = .79$; $p_{lightverb} = .27$; $p_{transparent} = .7$) and for familiarity ($p_{sense-selection} = .849$, $p_{transparent} = .926$, $p_{light-verb} = .905$).

The concreteness ratings for the resulting set of stimuli, which we will use later on in the encoding analyses as a way to capture the semantics of the phrases (see Section 4.1), are reported in Fig. 3. They are placed along a linear continuum, where it is possible to visually retrieve the distinction in concreteness between physical objects and informational contents. Note also that the position of a phrase along the gradient appears to be determined in part by the noun, in part by the verb, as shown by the fact that phrases with the same noun occupy different positions along the gradient.

The concreteness ratings in Fig. 3 differ significantly across phrases for physical objects and information, both overall ($p = 4.51e - 06$) and for each semantic composition case. As

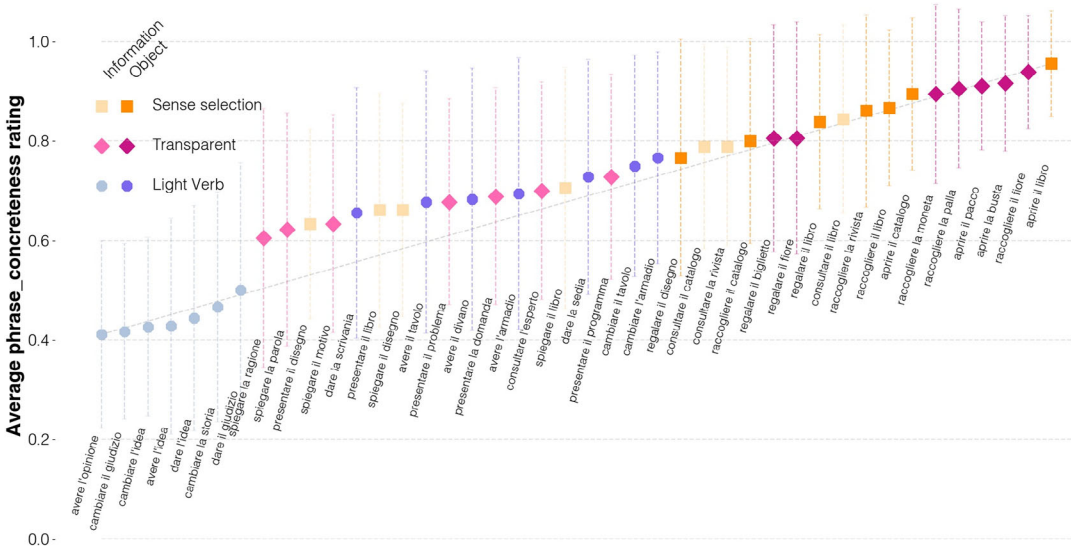

Fig. 3. Visualization of the average concreteness ratings for the 42 stimuli, provided by 36 raters. Each phrase's concreteness is represented by a marker whose color and shape also indicate which semantic type (physical object or information) and which composition case it belongs to. The position on the *y* axis of each point corresponds to the average concreteness value across our 36 raters; error bars represent the standard deviation. Below each marker, we report the original phrase in Italian. Ratings were rescaled in the range 0–1 in order to ease readability.

expected, a lower *p*-value is obtained for the types of composition where a clearer separation between abstract and concrete exists in principle, that is, the transparent composition phrases ($p_{transparent} = 8.69e-06$) and the light verbs phrases ($p_{light-verb} = 1.29e-07$), whereas sense selection phrases are more nuanced ($p_{sense-selection} = .009$): after all, they involve the same word across semantic types (e.g., *book* can be both a physical object and an informational content).

3.2. fMRI data

3.2.1. Participants

Nineteen volunteers were recruited, but three of them were excluded from the analysis because they failed to respond or respond incorrectly in more than 10% of the trials, leaving 16 participants for the analysis (8 females; average age 22.5, standard deviation 3.42). All participants were native Italian speakers, right-handed, and had normal or corrected-to-normal vision.

3.2.2. Data acquisition

All of the fMRI experiments were conducted with a 4T Bruker MedSpec MRI scanner. Structural images were acquired using a T1-weighted MPRAGE sequence with resolution 1*1*1mm. A T2*weighted EPI pulse sequence was used to acquire the functional images with parameters TR 1000 ms, TE 33 ms, and 26 flip angle, FoV1000*1000. Each acquisition

volume contains a 64*64 matrix and 17 slices with a gap of 1 mm. Voxel dimensions are 3mm*3mm*5mm.

3.2.3. Experimental paradigm

Participants were instructed to attentively read the phrases and judge whether the verb-noun combinations were meaningful. This sensicality task was already employed in previous brain studies on semantic composition (Pylkkänen, Llinás, & McElree, 2004; Pylkkänen, Oliveri, & Smart, 2009; Schell et al., 2017) and polysemy (Husband et al., 2011; Pylkkänen et al., 2006). About 10% of the stimuli were catch-trials with meaningless combinations (e.g., *open the sun*). Each trial started with a fixation cross for 500 ms, followed by a verb and then an article-noun phrase, where each was presented for 450 ms, with a 100-ms interval. A black cross then remained on the screen for 1500 ms, and subsequently, a question mark was displayed for 1000 ms, where participants had to respond whether the presented phrase was meaningful, by pressing the left or right button box (counterbalanced across participants). The next trial started after a fixation time of 6 s. During one scanning session, all 42 verb-noun phrases along with five catch-trials (10% of all trials) appeared once in a random order. Each participant completed six sessions.

3.2.4. Preprocessing

We preprocessed the fMRI data with SPM12 (Penny, Friston, Ashburner, Kiebel, & Nichols, 2011). We used default parameters for all steps, unless specified otherwise. First, we defaced the anatomical brain images, in order to guarantee anonymity; then, we realigned the images and corrected for the timing acquisition of the slices. Next, we coregistered the T1 to the mean EPI image; and finally, we normalized the images to the MNI space, keeping the original voxel size of 3mm*3mm*5mm. We did not smooth the images.

To obtain one BOLD response image for each phrase, capturing how the brain processes verb-noun semantic composition, we followed the methodology of Mitchell Anderson, Zinszer, and Raizada (2004, 2008), and Anderson, Zinszer, and Raizada (2016). Therefore, we averaged the BOLD response corresponding to the time points between 4 and 8 seconds (s) after the presentation of the noun, thus accounting for the delayed hemodynamic response to the stimulus.

It is important to stress that we used the presentation of the noun, instead of the verb, as the starting point (t_0) for the evoked BOLD response. This is due to the fact that, following previous work on semantic composition in fMRI and MEG (Bemis & Pylkkänen, 2011; Brennan & Pylkkänen, 2012; Husband et al., 2011; Zhang & Pylkkänen, 2015; Zaccarella et al., 2017), in our paradigm, the verb and the noun were presented one after the other (serial presentation) and not concurrently (parallel presentation; Snell & Grainger, 2017). This being the case, semantic composition could only take place after the appearance of the noun. We empirically validated our assumption that semantic processing would start only after the presentation of the noun in a time-resolved encoding analysis. We report the results in Appendix B in the Supplementary Materials (Figs. B.1– B.4). Time-resolved analyses allow to understand how each model captures semantic composition in the language network along the temporal dimension—and importantly, when semantic information starts to be present in brain activity.

Encoding was carried out separately for each TR falling in the time window between –2 and 12 s after the appearance of the first visual stimulus, the fixation cross (since TR = 1 s, we in fact evaluated encoding once per second). Results clearly indicate that all of the model representations capture brain processing in the time window between 4 and 8 s after the presentation of the noun—but not earlier, confirming our assumption. We believe that, had we used a parallel presentation paradigm, instead of the serial methodology we actually employed, results would not have differed significantly. Our expectation is that, as indicated in work on access to phrase- or sentence- level representations (Snell & Grainger, 2017), semantic processing would have simply taken place slightly earlier, given the immediate availability of an interpretation for the full phrase. This would have just shifted t_0 to the time of presentation of the full phrase, instead of the noun, as it was the case in our experiment.

3.2.5. Language network analysis

For the encoding analyses, where we compare the performances of a set of models, we focused on a specific brain network, the language network (Fedorenko et al., 2010). We employed, for our encoding analyses based on this network, the brain mask provided by Fedorenko et al., which consists of a bilateral map obtained from the aggregation of the results of the localizer task on 220 subjects.[3] When masked using the language network, our brain images are composed of a total of 2392 voxels.

3.2.6. Regions-of-interest analysis

For the ROI analyses, we focused on understanding which brain regions contain most information with respect to the process of semantic composition. Therefore, we used the model with the best performance—the phrase concreteness model—as the predictor for the encoding. We focused on five regions of interest (left IFG, left ATL, left pSTS, vMPFC, and motor areas) which have been previously implicated with semantic composition. To isolate the left ATL (295 voxels), the left IFG (182 voxels), and the left pSTS (637 voxels), we used the masks available within the manual parcellation of the language network provided by Fedorenko et al. (2010). For the vMPFC, we used the vMPFC mask published by Delgado et al. (2016) (727 voxels). For the motor areas, we used the masks for the precentral gyrus and the supplementary motor cortex available from the Harvard-Oxford Brain Atlas (Desikan et al., 2006) (1789 voxels).

3.2.7. Feature selection

Feature reduction is fundamental for fMRI data, which has high dimensionality. Dedicated methods have been devised in the literature for encoding studies like ours, where a mapping function is learnt between a vectorial model and brain data (Mitchell et al., 2008; Pereira et al., 2018). We adopted the methodology of Mitchell et al. (2008), which is a straightforward choice in encoding studies involving word vectors (Anderson, Zinszer, & Raizada, 2016; Caceres et al., 2017). This procedure was carried out separately for each subject separately and for each train-test split (see Section 5.2). Within each training set, first features (voxels) are ranked according to their so-called stability (i.e., average Pearson correlation of BOLD signal across the six trials for each stimulus, with higher correlation values indicating higher

stability). Then, the top n most stable voxels are selected and retained for further analysis; the same selection of features, determined independently from the test set, is applied to the test set. As in Mitchell et al. (2008) and Anderson et al. (2016), we used the $n = 500$ most stable voxels. Notice that the only difference of our approach with the original implementation is that voxels were not selected from the whole brain, but from the language network or each ROI, depending on the analysis. Feature selection was not applied for those brain areas whose total number of voxels was lower than 500 (left IFG and left ATL; see Section 3.2.6).

As shown in Fig. 1, where we report the 25% most used features plotted against the language network, the feature selection procedure selected voxels belonging mostly to the left pSTS, but also, in minor part, to the left ATL and the left IFG.

4. Models

4.1. A "cognitive model" of concreteness

The main semantic dimension along which our set of 42 phrases, which exemplify different modes of verb-noun semantic composition, vary is **concreteness**, since the nouns in the stimuli can refer to either a physical object or to information. Concreteness has been found to be a variable playing a fundamental role for semantic processing in both behavioral and brain studies (Mkrtychian et al., 2019; Montefinese, 2019). Early approaches viewed the abstract/concrete distinction as binary, and involving different processing pathways—for instance, the Dual Coding theory of Paivio (1969), or the interpretation of results from patients of Crutch and & Warrington (2005). However, a more graded view of the distinction between concrete and abstract semantic representations has emerged in more recent literature. In this literature, concreteness and abstractness are organized along a continuum, with gradual involvement of sensorimotor, linguistic, and emotional features (Anderson et al., 2014; Borghi et al., 2017; Glenberg et al., 2008; Ghio, Vaghi, & Tettamanti, 2013; Hill, Korhonen, & Bentz, 2014; Troche, Crutch, & Reilly, 2017).

Therefore, we also learned encoding mappings between a graded notion of concreteness and fMRI data. Being able to find an accurate mapping of this kind would provide evidence that concreteness is one of the dimensions affected by semantic composition, shifting word meanings toward more or less concrete interpretations (cf. Hill & Korhonen, 2014).

Also, we believe that a comparative approach, looking at the difference in performance between this model and a computational language model like ITGPT2, can provide interesting insights with respect to the ability of current language models to capture subtle features of semantic composition in the brain. In this sense, we expect concreteness to provide an upper bound on the quality of our language models in terms of encoding, since concreteness ratings are provided by human subjects. We nevertheless acknowledge that the types of semantic information contained in human ratings and language models differ along many dimensions, with language models encoding disparate types of semantic information at once (Emmanuele, Enrico, Chu-Ren, Lenci, & others, 2021; Utsumi, 2020). Evidence emerging from the contrast of the two, therefore, should not be interpreted in terms of a strict alignment of the two types of

representation, but in terms of a broad convergence toward representations that can similarly explain brain activity (Blank, 2023; Schrimpf et al., 2021).

4.1.1. Phrase concreteness

As a model of complex composition (see Introduction) based on concreteness, we used the concreteness scores obtained through the rating procedure presented in Section 3.1, that we call **phrase concreteness**. In this case, we assumed, following Naselaris et al. (2011) that, when confronted with the task of rating the concreteness of the full phrase, raters are in fact carrying out a nonlinear transformation of the lexical representations into a complex representation of their composed meaning, which they then express in terms of a phrase concreteness rating.

4.1.2. Single-word concreteness

We wanted to create a model of simple composition (see Introduction) based on concreteness. The fundamental constraint of such a model is that the representation of a phrase should be obtained without modifying the representations of the individual concepts involved in complex, nonlinear ways. To achieve this, we used a methodology which has been used in previous work as a baseline to capture how linguistic context shapes semantic dimensions such as concreteness (Gregori et al., 2020) or valence, dominance and arousal (Calvo & Mac Kim, 2013). This approach relies on averaging the ratings for individual words contained in a phrase to obtain a representation of the phrase's relevant rating. For instance, suppose the words "imagine" and "cello" are, respectively, rated $concreteness_{imagine} = 2$ and $concreteness_{cello} = 5$. The rating for the phrase "imagine the cello" would be $concreteness_{imaginethecello} = (2 + 5)/2 = 3.5$. Notice that, despite its extreme simplicity, averaging has proven to be a strong mechanism to capture facets of meaning not only in computational modeling (Dinu, Baroni, & others, 2013) but also in brain studies (Wu, Anderson, Jacobs, & Raizada, 2022). As a starting point, we collected a set of ratings from a group of Italian raters ($n = 36$) for the individual words (verbs and nouns) contained in the phrases. Then, we modeled each phrase's concreteness as the average of the concreteness of the verb and the noun. We call this model **single-words concreteness**. The phrase concreteness model and the single-words concreteness models present a moderate degree of representational similarity ($r = .672$, Fig. 4), confirming that they capture different, but related approaches to the representation of the semantics of the phrases.

4.2. Language models

4.2.1. Modeling compositionality with distributional semantics and language models

Much work has been dedicated, in distributional semantics, to the topic of compositionality (for overviews, see Baroni et al., 2014; Erk, 2012; and Mitchell & Lapata, 2010). Compositional distributional semantics as a research field is particularly active for *static* language models. This is because static models represent individual words in isolation: if one wants to capture the subtleties of meaning composition as it is carried out by humans, it is an open question how to best compose such individual vector representations (Boleda, 2020).

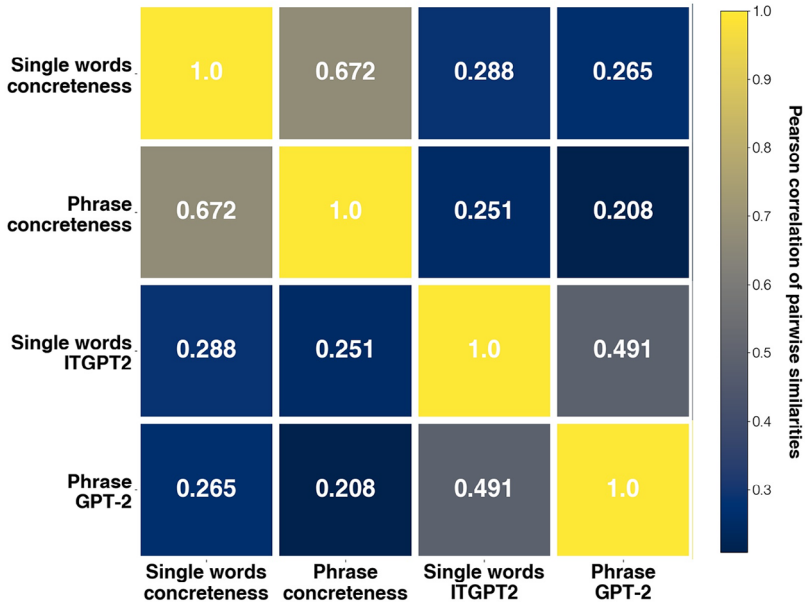

Fig. 4. Representational similarity among the models' representations used for encoding. The color and the value reported in each cell of the confusion matrix reflect the Pearson correlation among the corresponding row and column models' representational spaces, computed as is customarily done in representational similarity analysis, as the vector of within-model pairwise Pearson correlations (Diedrichsen & Kriegeskorte, 2017). Overall, models seem to capture in distinct ways the semantics of the phrases, as is shown by the gradient of correlations: in particular, concreteness models are more similar to one another than computational models.

In other words, the core challenge for compositional distributional semantics approaches is being able, given a linguistic context such as a sentence, to infuse knowledge about syntactic and semantic structure in vectors for individual words. Vanilla distributional word vectors bear no explicit trace of syntax, or part of speech information, or semantic category, aside from what can be indirectly captured through distributional information. In other words, the vectors for words like "explain," "music," and "the" (respectively, a verb, a noun, an article) all live in the same vector space, without any clear boundary between them, as is instead the case, for instance, in grammars (notice also that the different senses of "book" are squeezed in a single vector). This limits the ability of such models to capture fine-grained semantic composition, which relies on various pieces of information emerging at the interface between syntax and semantics. Various solutions have been proposed in the literature, mostly based on tensors or special composition operations (Baroni & Lenci, 2010; Baroni & Zamparelli, 2010; Baroni, 2013; Chersoni et al., 2019; Erk & Padó, 2008; Grefenstette, Dinu, Zhang, Sadrzadeh, & Baroni, 2013; Lenci, 2011; Levy & Goldberg, 2014). Contextualized language models, on the contrary, are designed to overcome such limitations by default. They represent words *in context*. Therefore, they specialize in capturing the way in which linguistic meaning is shaped compositionally in context. To do so, they heavily rely on the use of machine learning mechanisms like attention and self-attention (Vaswani et al., 2017). Because of this

they constitute the best models currently available from NLP when it comes to capturing fine-grained variation in meaning triggered by compositional processes (Apidianaki, 2022). This drove our decision to use a contextualized language model for our analyses.

4.2.2. ITGPT2

We tested two separate types of representations for the contextualized language model—modeling a complex and a simple approach—mirroring the representations obtained for concreteness. Starting from the same model, we followed two different vector extraction procedures—one for complex composition, and another one for simple composition (described below). We would like to underline that, for both simple and complex semantic compositions, we used the exact same pretrained language model. This allowed us to eschew any possible confound due to pretraining factors, like size and source of training data used, or the number of parameters in the neural network. Such variables differ dramatically across language models and strongly affect their performance, making it hard to make reliable comparisons (Kaplan et al., 2020; Kirstain, Lewis, Riedel, & Levy, 2022; Min et al., 2021; Zhang, Warstadt, Li, & Bowman, 2021). As a starting point for both the simple and the complex models of composition, we used ITGPT2 (de Vries & Nissim, 2021), an adaptation for Italian of GPT2, a model which is publicly available and widely used in the literature (Radford et al., 2019). Our choice of ITGPT2 was guided by the fact that its English counterpart, GPT2 (Radford et al., 2019), had been previously used and validated with brain data for English (Caucheteux et al., 2021; Schrimpf et al., 2021). Models in the GPT family are available in different sizes, depending on the number of parameters used in the neural network. As a rule of thumb, more parameters certainly increase the computational cost—sometimes making a model impossible to use without dedicated computing resources (Izsak, Berchansky, & Levy, 2021)—but may or may not improve performance (de Varda & Marelli, 2023; Kaplan et al., 2020). To obtain the best results possible, we use the best-performing version of ITGPT2 available, which is ITGPT2-medium. This deep neural network has 24 layers and 380 millions of parameters, just like its English equivalent, GPT2. ITGPT2 was created using a two-steps procedure. First, the input (embedding) layer of the original GPT2-small English model (124 millions of parameters) was retrained, leaving the rest of the neural network frozen, for Italian, using ItWaC and Wikipedia in Italian (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). This effectively operates a translation of the input layer embeddings of the pretrained English model. Then, the newly learnt input layer embeddings were mapped to the bigger GPT2-medium model (having 355 millions of parameters) using a linear transformation. This led to the creation of a version of GPT2-medium "translated" to Italian. For reference, the original GPT2 English model was trained on 40GB of text scraped online, from which Wikipedia pages were removed (Radford et al., 2019). ITGPT2 inherits the knowledge extracted by GPT2 from such training data in its hidden layers, above the input layer which is instead adapted to Italian. To validate our choice of using ITGPT2 as our contextualized model, in Appendix D in the Supplementary Materials, we report the results of comparing its performance against various versions of XGLM (Lin et al., 2021), a much bigger language model (1.7, 2.9, 4.5, and 7.5 billions parameters). This model was created as a publicly available counterpart to GPT-3, and was trained on CC100-XL, a huge multilingual corpus covering 68 snapshots of the Common

Crawl dataset and 30 languages, including Italian (Lin et al., 2021). Results indicate that ITGPT2 is always better than all of the versions of XGLM, no matter their size (Appendix D, Fig. D.1; see Section 7.2 for a discussion on this point which may seem counterintuitive). In Appendix E in the Supplementary Materials, we also validate empirically our assumption that a contextualized language model would perform better than a static counterpart (Appendix E, Fig. E.1). We report the encoding results using FastText, a state-of-the-art static language model (Bojanowski et al., 2017), in a variety of vector extraction modalities—in all cases, results are always worse than those obtained with either XGLM or ITGPT2.

4.2.3. Vector extraction procedure

To extract the contextualized vectors to be used for encoding, we adapted the so-called procedure of "representation pooling," a methodology validated on benchmarks from computational linguistics in Bommasani et al. (2020), Vulić et al. (2020), and Apidianaki (2022) and, more specifically, in Bruera and & Poesio (2022), for brain data. The procedure is equal for both single-words ITGPT2 and phrase ITGPT2. The differences between the two models will be explained in detail below. For the time being, it will be enough to say that, for single-words ITGPT2, we extracted representations for words in isolation (e.g., "open," "book"); for phrase ITGPT2, by contrast, we extracted the representations for the phrases directly (e.g., "open the/a book," in the context of a sentence). We implemented the representation pooling vector extraction procedure in five steps:

1. collecting from natural language corpora sentences containing mentions of each phrase (phrase ITGPT2) or either the verb or the noun (single-words ITGPT2). For phrase ITGPT2, we sampled sentences where the verb preceded the noun by no more than two words, so as to be able to capture cases like "open the old book." For single-words ITGPT2, we sampled the mentions independently for verbs and nouns, so as to capture contexts of usage of the individual words which were not related to the phrases themselves (all the sentences used are publicly available together with the code and the extracted vectors);
2. encoding the sentences using the contextualized model;
3. extracting the top 12 hidden layers (which in Jat et al., 2019; Schrimpf et al., 2021; Antonello, Vaidya, & Huth, 2023 have been shown to best capture brain semantic processing) for the tokens corresponding to the phrase (phrase ITGPT2) or word (single-words ITGPT2). We also consider as part of the phrase representation the hidden activations for the first token occurring after the phrase, as we found that this had a positive impact on results. This choice is due to the fact that GPT2 is a causal language model—that is, it learns to predict the coming word in a sentence given a previous sequence of words. In these models, the hidden states at $t + 1$ capture the information contained in the whole sequence until t. Therefore, the hidden representation at $t + 1$ contains information about the whole phrase that came before;
4. for each mention, averaging across layers and tokens, so as to obtain a single mention contextualized vector;

5. finally, averaging across at most n randomly selected mentions of each phrase (phrase ITGPT2) or word (single-words ITGPT2). Following Vulić et al. (2020), $n = 10$ since this amount of vectors has been shown to provide optimal results.

This methodology provided us with one contextual vector for each phrase (phrase ITGPT2) or word (single-words ITGPT2). For phrase ITGPT2, this captured an averaged meaning of the phrase in various linguistic contexts. For single-words ITGPT2, an additional step was required (mirroring the procedure followed for phrase concreteness). We will describe it below.

As corpora to extract the sentences from, we used three different types of texts in Italian, in order to maximize the coverage of our contextual phrase vectors. The first was the Italian version of Wikipedia, which is commonly used in the creation of distributional word vectors (Bojanowski et al., 2017; Bommasani et al., 2020; Devlin et al., 2019). The second was the Italian portion of the OpenSubs corpus of film subtitles (Tiedemann, 2012), since it has been shown that subtitles allow for the creation of word vectors that can model psycholinguistic phenomena better than generic written corpora (Mandera, Keuleers, & Brysbaert, 2017). The third was ItWac, a corpus of Italian texts crawled from the Web (Baroni et al., 2009), which is again a common choice in the field (Levy, Goldberg, & Dagan, 2015; Mandera et al., 2017).

We report in Appendix C in the Supplementary Materials an ablation study for ITGPT2, focusing on how each the style of each corpus affects the vectorial representations for the phrases (Appendix C, Fig. C.1). In it, we investigated the impact on encoding results of removing sentences from one of the three corpora before representation pooling for the full-phrase ITGPT2 (in fact, it was not possible to remove ItWac, as it was the only corpus containing at least one example per phrase). As it can be seen, using all three corpora led to superior performance overall. Especially for the sense selection and light verb cases, using all three corpora resulted in a clear improvement in encoding results. This validates our approach, where the full set of corpora is used, and indicates that a mixture of styles seems to be beneficial for representation pooling to be used in brain encoding.

In Fig. 4, we report the representational similarity between representations from ITGPT2 and the concreteness models, indicating that they seem to capture rather different types of semantic information. Full-phrase ITGPT2 correlates about the same with both concreteness models ($r = .208$ for phrase concreteness, and $r = .265$ for single-words concreteness). Single-words ITGPT2 is more similar to concreteness models than phrase ITGPT2 ($r = .288$ for single-words concreteness, and $r = .251$ for phrase concreteness). Finally, correlations between simple and complex models are moderate, indicating relevant differences among the two types of representations (ITGPT2: $r = .491$; concreteness: $r = .672$).

4.2.4. Phrase ITGPT2

For phrase ITGPT2, the representations for each phrase are the ones obtained at the end of the vector extraction procedure described in Section 4.2.3.

4.2.5. Single-words ITGPT2

For single-words ITGPT2, we used a methodology directly matched to the one used for single-words concreteness, the cognitive model of simple composition. Through vector

extraction (see Section 4.2.3), we obtained separate semantic representations for individual verbs and nouns, just as was done for concreteness ratings (see Section 4.1.2). The representations for the phrases were then composed by averaging the vectors for the verb and the noun. This counts, as discussed above, as a simple method of meaning composition. Here, the composition process does not alter in nonlinear ways the meaning of the parts being composed, as is instead the case for phrase ITGPT2. It is important to underline that, for both phrase ITGPT2 and single words ITGPT2, the vectors come from the same deep layers of the same neural network. The crucial difference is that while the vectors for the phrases in phrase ITGPT2 are contextualized *phrase vectors*, the representations of the phrases in single-words ITGPT2 are averages of contextualized *individual word* vectors. This mirrors as closely as possible the difference between the two cognitive models, phrase concreteness and single-words concreteness (see Section 4.1.2).

4.3. Signal-to-noise ratio

We also report so-called **ceiling** encoding values (Anderson et al., 2019, 2021; Nili et al., 2014), which are a way to quantify the signal-to-noise ratio (SNR) in the fMRI data. The ceiling value indicates to what extent the fMRI data in itself can be encoded—how distinguishable in the fMRI data are the patterns corresponding to the stimuli (Anderson et al., 2019). The ceiling values were computed using exactly the same encoding setup (see Section 5.2), with the exception that, as inputs, averaged fMRI responses from other subjects (without any feature reduction) are employed instead of model representations.

Notice that, since ceiling values are obtained from mapping brain recordings between subjects, they are in fact themselves "noisy" measurements. This is due, first, to the noise inherent in the measurements, as well as noise related to inter-subject alignment—and we assume that, increasing the sample size, estimates would be more precise (Cremers, Wager, & Yarkoni, 2017; Desmond & Glover, 2002). Second, additional noise comes from the fact that it is not defined exactly what type of information is contained in ceiling scores (Anderson et al., 2019). This means that they should not be considered as absolute upper bounds on performance—in fact, top-performing models can provide performance above the SNR ceiling (Anderson et al., 2019; Schrimpf et al., 2021). They should simply be interpreted as a guide for the interpretation of the results: SNR ceiling values, in our case, provide an approximate expectation of what level of performance a good model should be expected to achieve.

Ceiling values are reported in the plots as gray shades (the ceiling value for each condition is the liminal value where the gray shade disappears).

5. Brain encoding

5.1. Methodology

To find the mapping between brain data and models, we used the representational similarity encoding framework of Anderson et al. (2016). Intuitively, this approach does not require to

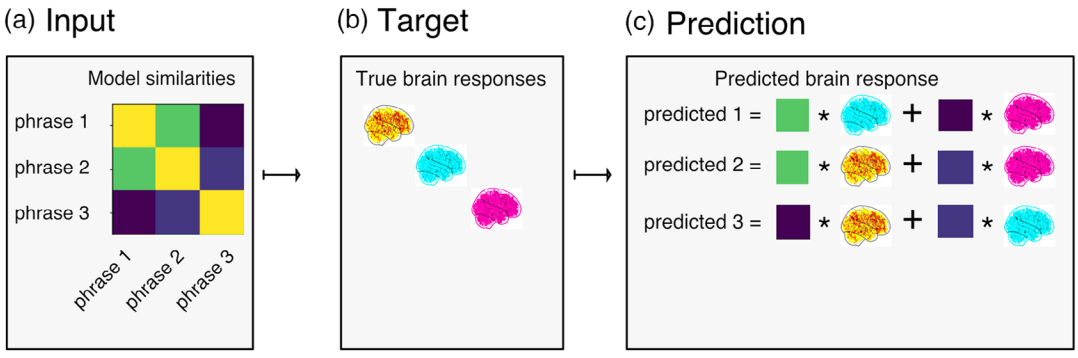

Fig. 5. Visualization of the RSA encoding procedure using a toy example with only three phrases. Here, we consider a simple example, where we want to encode three brain images. In the RSA encoding framework, there is no model to be fit to make the predictions, and only similarities in a given model are used. Starting from the input (part A), which are similarities between phrases computed using a model (in our case, Pearson correlation for language models and inverse of the absolute difference for concreteness ratings), the prediction for each phrase (part C) is obtained as the sum of the real brain images for the words outside of the test set (in our example, the brain images for phrase 2 and phrase 3 from part B), where each image is weighted by the similarity in the model with the target phrase. Then, when all target images have been predicted, the leave-two-out pairwise evaluation described in Section 5.2 takes place. Notice that, since in the actual evaluation each test set is composed of two test items, not of one single item as is the case in this toy example, in our experiment when predicting the brain images (part C), similarity to the other test item is not used to make the prediction.

fit a model: instead, it simply relies on pairwise similarities among brain images and model representations in their native spaces to predict a brain image (encoding)— details are provided below.

This method, despite its simplicity, provides multiple advantages: excellent performance, as shown in Anderson et al. (2016); straightforward interpretability in the framework of the representational similarity analysis (RSA: Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013); no risk of overfitting (Hosseini et al., 2020).

5.1.1. Representational Similarity Encoding

In encoding, the goal is to predict the brain response to a stimulus given its model representation. In representational similarity encoding (Anderson et al., 2016), the brain image for a given stimulus from the test set is modeled as the weighted sum of the brain responses to the stimuli in the training set, where the weights are the pairwise Pearson correlation between the test item and each training item. For instance, given a toy training set of three model representation for the phrases $\vec{a} = \textit{read the book}$, $\vec{b} = \textit{throw the book}$, $\vec{c} = \textit{copy the book}$ and their corresponding brain images a_{brain}, b_{brain}, c_{brain}, the brain response d_{brain} to the test item $d = \textit{open the book}$ given its model representation $\vec{d}$ would be computed as $d_{brain} = a_{brain} * r_{\vec{a}*\vec{d}} + b_{brain} * r_{\vec{b}*\vec{d}} + c_{brain} * r_{\vec{c}*\vec{d}}$ where r is the operation of Pearson correlation (see Fig. 5 for a visualization of this procedure).

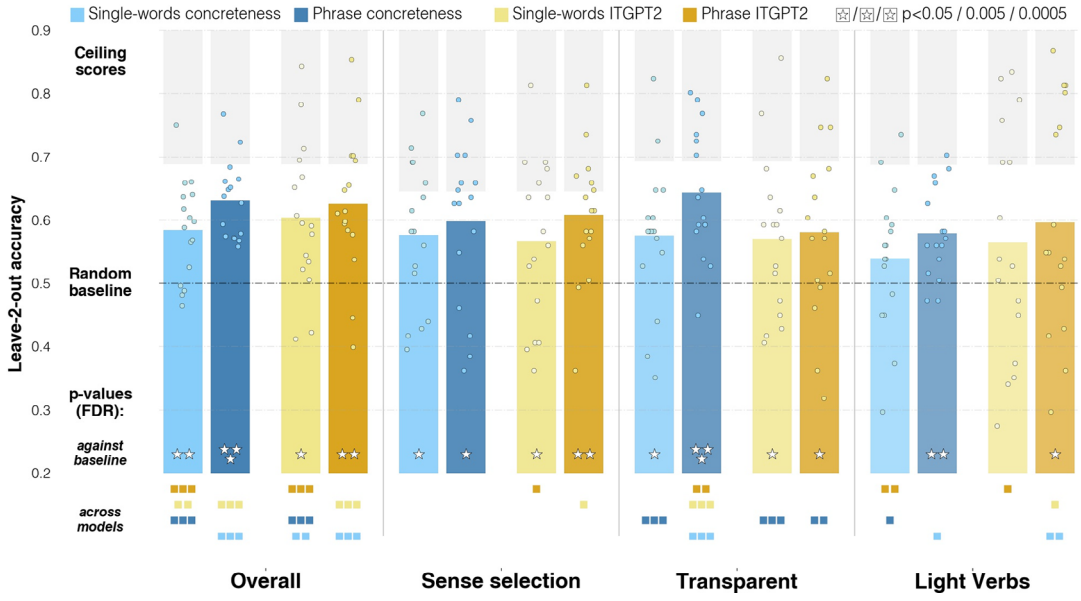

Fig. 6. Comparing encoding scores across models and composition modes. Left to right: accuracy of all stimuli (overall) and the three modes of composition (see Section 5.2). The two cognitive models are in shades of blue and the two language models are in shades of yellow. Complex models are in the respective darker shade. The *Y*-axis represents the accuracy score, averaged across subjects. Bar heights correspond to average values, and average scores for individual subjects are reported as dots. Ceiling values, which quantify the SNR—that is, the level of best possible encoding or decoding of the fMRI data itself, are reported as inverted gray bars, going from 1. to the ceiling accuracy value. The random baseline of 0.5 is indicated by a dotted line. We report the results of statistical tests against the baseline as stars in the lower part of each bar (at a y value of slightly above 0.2); one star stands for $p < .05$, two stars for $p < .005$, and three stars for $p < .0005$. Below the plots, we report pairwise statistical comparisons for each possible pair of bars within each section of the plots, using squares whose color reflect the model whose comparison is reported. All p-values are FDR-corrected (see Section 5.3). Complex models of composition are always above chance with statistical significance, and better than simple models. The difference between simple and complex models is also statistically significant in most cases.

5.2. Evaluation

In this study, we used leave-two-out pairwise evaluation, commonly used in brain encoding studies involving vectorial computational models (Honari-Jahromi et al., 2021; Mitchell et al., 2008; Pereira et al., 2018). This methodology is also suited to the representational similarity encoding framework, as it was the one used in Anderson et al. (2016). This procedure is repeated for all possible pairs of stimuli within each subject. At each iteration, the data are split into a train set, comprising all but two stimuli, and a test set, which instead is made up by the two left-out stimuli. From the training set, predictions for the two left-out stimuli are produced. The accuracy for an iteration is evaluated by computing the four Pearson correlations between the predicted and target vectors, and comparing the "matched" and "mismatched" scores to obtain a binary evaluation score for the current test iteration, where 1 equals correct decoding and 0 is for wrong decoding. Therefore,

$accuracy = 1$, if $corr(\vec{target}_1, \hat{\vec{target}}_1) + corr(\vec{target}_2, \hat{\vec{target}}_2) > corr(\vec{target}_1, \hat{\vec{target}}_2) + corr(\vec{target}_2, \hat{\vec{target}}_1)$; else, encoding is considered unsuccessful, and $accuracy = 0$. The assumption is that for the model to have learnt how to carry out the encoding means that the overall correlation among correctly matched vectors should be higher than those for the incorrectly matched vectors (Mitchell et al., 2008).

Scores across all iterations are then averaged within a subject. An average accuracy of 0.5 indicates chance performance, since the scores are binary. Encoding analyses are carried out at the level of individual subjects, and then averaged across all subjects to provide the final overall encoding score.

Note that, each round of pairwise evaluations measures to what extent the model has learnt to generalize in the specific case of the categorical relation holding between the two test items (Bruera & Poesio, 2022; Chyzhyk, Varoquaux, Milham, & Thirion, 2022; Elangovan, He, & Verspoor, 2021; Grootswagers et al., 2017; Gorman & Bedrick, 2019; Lake & Baroni, 2018). Therefore, it is possible to examine the results of different categorical relations separately. For instance, we can examine to what extent the model captures semantic composition involving light-verbs by looking at the accuracy of the iteration rounds that consist of phrases belonging to the light-verb composition mode as the test stimuli (e.g., *have the envelope* vs. *have the flower*).

5.3. Statistical testing and multiple comparisons correction

We ran two sets of statistical significance comparisons. First, we measured whether encoding accuracies were reliably above chance ($chance = 0.5$) with one sample t-tests. Second, we compared the results for each pair of representational models with one another and of each pair of ROIs. For this, we used McNemar's test, which is the standard way of comparing binary scores produced by a machine learning model (Stkapor, 2017). p-Values were corrected with the False Discovery Rate (FDR) procedure for multiple comparisons (Benjamini & Hochberg, 1995). We correct scores for multiple comparisons separately for encoding and decoding and for each family of tests (t-tests and McNemar tests). Within a set of tests (e.g., t-tests), all p-values are corrected using only one procedure (i.e., p-values for all models/ROIs are concatenated and corrected using only one call of the function $mne.stats.fdr\_correction$). This ensures reliability of the correction procedures, avoiding incorrect rejection of the null hypothesis.

6. Results

6.1. Encoding semantic composition in the language network

Phrase concreteness gives the best encoding scores for all composition cases, with strong statistical significance (overall= 0.63, $p < .0005$, sense selection = 0.598, $p = .0105$, transparent= 0.643, $p = .0002$, light verbs= 0.579, $p = .0014$). The single-words concreteness model, despite often having statistically significant differences to the phrase concreteness model ($p_{overall} < .0005$, $p_{sense\ selection} = .4089$, $p_{transparent} < .0005$, $p_{light\ verbs} = .0479$),

nevertheless reaches statistical significance in almost all composition cases (overall= 0.584, $p = .0016$, sense selection= 0.576, $p = .0235$, transparent= 0.575, $p = .0253$, light-verbs= 0.539, $p = .1154$). This indicates that modeling verb-noun semantic composition in terms of a graded concreteness space captures a lot of the signal present in the brain.

The performance of the contextualized language model, ITGPT2, shows the same pattern of results as concreteness, although differences are slightly less pronounced. Phrase ITGPT2 performs significantly better than chance across all cases (overall= 0.625, $p = .0016$, sense selection= 0.608, $p = .0018$, transparent= 0.581, $p = .0347$, light verbs= 0.596, $p = .0406$). In contrast, single-words ITGPT2 achieves statistical significance in all cases except light verbs (overall= 0.604, $p = .0061$, sense selection= 0.566, $p = .0481$, transparent= 0.57, $p = .0388$, light verbs= 0.564, $p = .1154$). Phrase ITGPT2 outperforms single-words ITGPT2 in all cases, and the difference between the two models reaches significance in all cases except transparent composition ($p_{overall} < .0005$, $p_{sense\ selection} = .0110$, $p_{transparent} = .7157$, $p_{light\ verbs} = .0169$). This suggests that, in the case of ITGPT2, the advantages provided by adapting the semantic representation of the lexical items to the context—the key feature which is exploited by phrase ITGPT2—allow us to capture quite well the nuances of fine-grained semantic composition in the brain.

This picture is also confirmed by direct comparisons between the two complex models of composition (phrase concreteness rating and phrase ITGPT2). Overall, phrase concreteness shows better performance than phrase ITGPT2, but the difference is statistically significant only for the transparent composition case ($p_{overall} = .7135$, $p_{sense\ selection} = .9025$, $p_{transparent} = .0006$, $p_{light\ verbs} = .6802$). Phrase ITGPT2, by contrast, appears to capture better than full-phrase concreteness ratings brain processing of verb-noun phrases containing light verbs and, marginally, sense selection—however, the differences between the two models, as reported above, do not reach statistical significance.

When looking at the simple models of composition, single-words concreteness is on a par with single-words ITGPT2 for the sense selection and transparent cases ($p_{sense\ selection} = .9025$, $p_{transparent} < 1.$), while the language model-based representations provide better encoding scores for the overall ($p_{overall} = .0008$) and light verbs ($p_{light\ verbs} = .3566$) cases.

In summary, the encoding results show that overall the "complex" approach to semantic composition—using a separate model for the entire phrase instead of computing its interpretation from that of the words—explains brain processing of the stimuli better than simple models. This pattern of results emerges primarily from the cognitive models, and is confirmed, although with less strong effects, by computational language models.

6.2. Encoding polysemy in the language network

We also checked whether we could distinguish the concrete and the abstract senses of each polysemous dot-object noun such as *book* (Fig. 7). We looked specifically at the subset of all pairwise results where different senses of the same noun were left out as items within the test sets (e.g., *book* as a physical object, as in *open the book*, and book as its informational content, as in *explain the book*). The phrase concreteness model reaches statistically significant encoding both for *book* and *magazine* ($book = 0.652$, $p = .0221$, $acc_{magazine} = 0.875$, $p = .0014$).

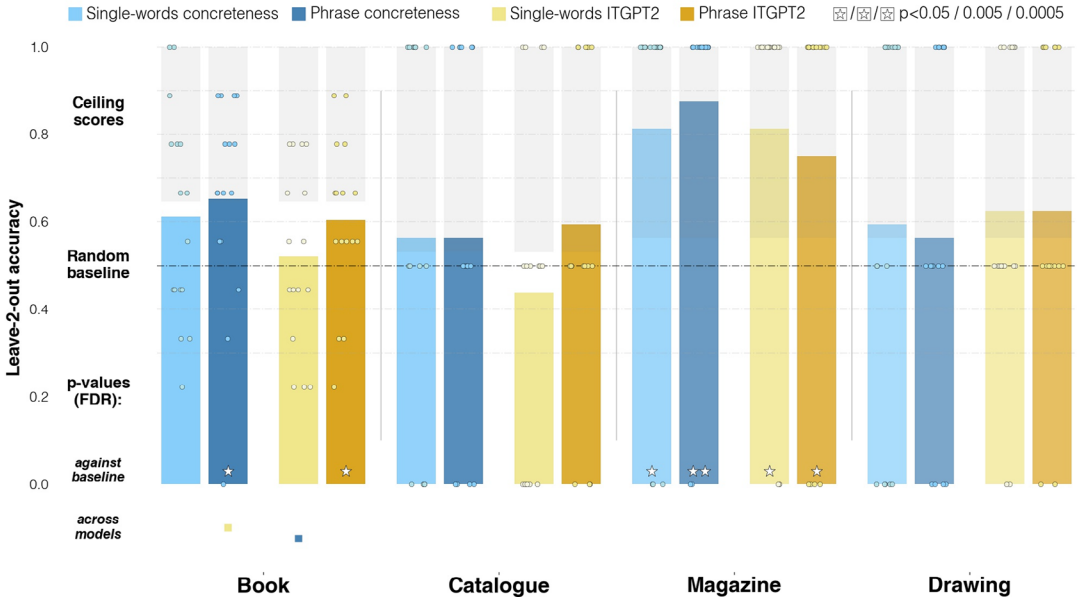

Fig. 7. Comparing encoding scores across models for the sense-selection mode. In the sense-selection mode, results for the polysemic nouns are reported separately along the *x*-axis. Left to right: accuracy of the four polysemous dot-object nouns. The two cognitive models are in shades of blue and the two language models are in shades of yellow. Complex models are in the respective darker shade. The *Y*-axis represents the accuracy score, averaged across subjects. Bar heights correspond to average values, and average scores for individual subjects are reported as dots. Ceiling values, which quantify the SNR—that is, the level of best possible encoding of the fMRI data itself, are reported as inverted gray bars, going from 1. to the ceiling accuracy value. The random baseline of 0.5 is indicated by a dotted line. We report the results of statistical tests against the baseline as stars in the lower part of each bar (at a *y* value of slightly above 0.2); one star stands for $p < .05$, two stars for $p < .005$, and three stars for $p < .0005$. Below the plots, we report pairwise statistical comparisons for each possible pair of bars within each section of the plots, using squares whose color reflect the model whose comparison is reported. All *p*-values are FDR-corrected (see Section 5.3). For the two senses of *book*, phrase concreteness provides the best scores, followed by phrase ITGPT2—both reaching significantly above-chance performance. *Magazine* is the only polyseme for which the two senses can be discriminated by all models.

The computational model of complex composition, phrase ITGPT2, shows largely similar performance ($book_{\text{ITGPT2}} = 0.604$, $p = .0396$, $magazine_{\text{ITGPT2}} = 0.75$, $p = .0388$). Simple models of composition, by contrast, perform worse overall: single-words concreteness performs well, but only approaches significance for *book* ($book_{single-wordsconcreteness} = 0.611$, $p = .0638$, $magazine_{single-wordsconcreteness} = 0.812$, $p = .0097$), while single-words ITGPT2 shows above chance performance just for *magazine* ($magazine_{single-wordsITGPT2} = 0.812$, $p = .0116$). Note that these results need to be taken with caution given the limited amount of data. This is particularly true for *catalogue*, *magazine*, and *drawing*, for which there are only two phrases each. However, for *book*, where six phrases are available, we did achieve better performance, and the pattern is similar to the sense selection results (reported in the first section to the left in Fig. 6).

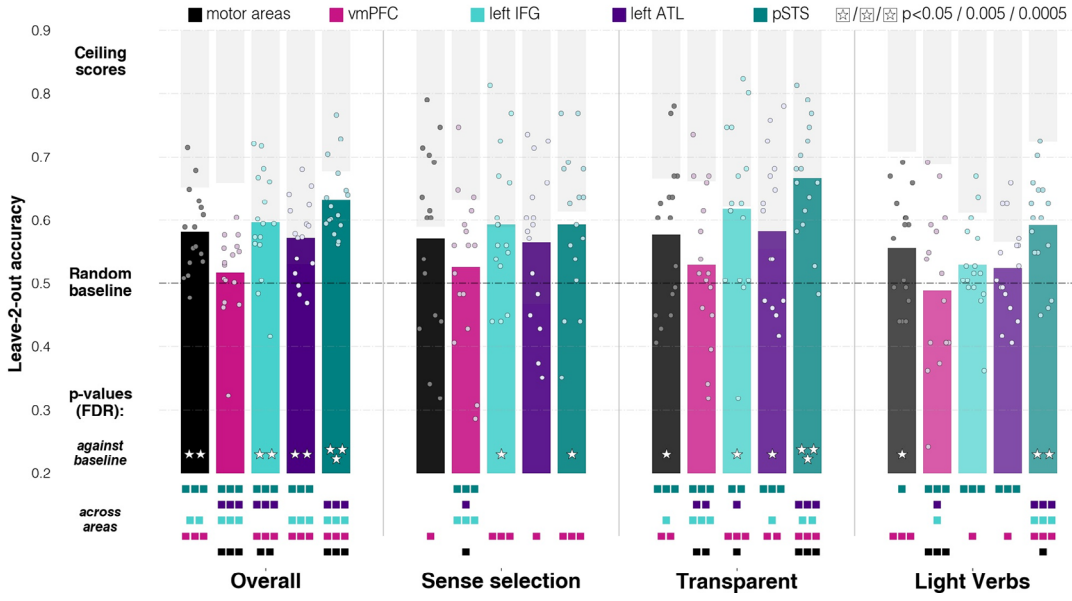

Fig. 8. Results for encoding phrase concreteness ratings from fMRI across the five regions-of-interest (ROIs). The *Y*-axis represents the accuracy score, averaged across subjects. Bar heights correspond to average values, and average scores for individual subjects are reported as dots. Ceiling values, which quantify the SNR—that is, the level of best possible encoding or decoding of the fMRI data itself, are reported as inverted gray bars, going from 1. to the ceiling accuracy value. The random baseline of 0.5 is indicated by a dotted line. We report the results of statistical tests against the baseline as stars in the lower part of each bar (at a y value of slightly above 0.2); one star stands for $p < .05$, two stars for $p < .005$, and three stars for $p < .0005$. Below the plots, we report pairwise statistical comparisons for each possible pair of bars within each section of the plots, using squares whose color reflect the model whose comparison is reported. All p-values are FDR-corrected (see Section 5.3). The left pSTS is the only brain area where encoding from phrase concreteness is statistically significant for all composition cases; the left IFG and the motor areas do so for the overall, the sense selection, and the transparent composition cases.

6.3. *Encoding semantic composition in individual brain areas*

We used a separate encoding analysis to find out in which ROIs previously implicated with semantic composition phrase concreteness could better predict brain activity related to verb-noun semantic composition. In this case, we chose to run the encoding using the phrase concreteness model only, since it is the model that theoretically captures most naturally the effects of semantic composition in our stimuli (see Section 4.1). Results are shown in Fig. 8.

The left pSTS (green in Fig. 8), the brain area which provides most of the voxels retained after feature selection (see Fig. 1), showed significant accuracy across all cases (overall= 0.632, $p < .0005$; sense selection= 0.593, $p = .013$; transparent= 0.666, $p < .0005$; light verbs= 0.592, $p = .0013$). Importantly, this is the only brain area where activity evoked by light verb cases can be predicted with statistical significance; in this case, difference with

other brain areas is strongly statistically significant too ($p_{light\ verbs} < .0005$ for the vMPFC, the left IFG, and the left ATL, and $p_{light\ verbs} = .014$ for motor areas).

The left IFG (cyan in Fig. 8) shows overall good encoding performance, with statistically significant scores in all cases except light verbs (overall= 0.596, $p = .0012$; sense selection= 0.593, $p = .0096$; transparent= 0.618, $p = .0066$; light verbs= 0.529, $p = .0991$).

The left ATL (purple in Fig. 8) shows lower performance than the left pSTS and the IFG, but encoding is still significantly above chance—or very close to significance—in all cases except light verbs (overall= 0.572, $p = .0012$; sense selection= 0.564, $p = .054$; transparent= 0.583, $p = .0233$; light verbs= 0.524, $p = .1572$).

Outside of the language network, the vMPFC (pink in Fig. 8) shows the lowest average performances and never reaches significance (overall= 0.516, $p = .2246$; sense selection= 0.526, $p = .265$; transparent= 0.529, $p = .2302$; light verbs= 0.489, $p = .6691$).

The most surprising results, however, come from the motor areas (black in Fig. 8). Despite being numerically lower in the left pSTS and the left IFG, performance is significantly above chance for all composition cases except sense selection, where it approaches it anyways (overall= 0.582, $p = .001$; sense selection= 0.571, $p = .0668$; transparent= 0.577, $p = .0287$; light verbs= 0.556, $p = .0288$). This indicates clear involvement in processing verb-noun semantic composition.

6.4. Encoding polysemy in individual brain areas

As in the language network analyses (Section 6.2), we also looked at encoding performance for the senses of each polysemous dot-object noun. This was done to evaluate where in the brain the concrete and the abstract senses of the dot-object nouns could be distinguished. The results are similar to the ones reported in Section 6.2: phrase concreteness can reliably predict brain activity related to polysemy with different patterns across brain areas. The two senses of *book* can be predicted with performance significantly above chance only in the left IFG ($book_{leftIFG} = 0.625$, $p = .0482$), whereas statistical significance is only approached in all other areas ($book_{leftATL} = 0.625$, $p = .054$, $book_{leftpSTS} = 0.618$, $p = .0814$, $book_{VMPFC} = 0.59$, $p = .0836$, $book_{motor-areas} = 0.59$, $p = .1022$). For the word *magazine*, statistically significant performance is found both inside and outside of the language network ($magazine_{VMPFC} = 0.875$, $p = .0015$, $magazine_{motor-areas} = 0.812$, $p = .0132$, $magazine_{leftATL} = 0.75$, $p = .0481$, $magazine_{leftpSTS} = 0.75$, $p = .048$; the only exception is $magazine_{leftIFG} = 0.687$, $p = .1022$). Finally, for *catalogue* and *drawing*, just as was the case in the analyses on the full language network (Fig. 7), encoding is never statistically significant ($catalogue_{VMPFC} = 0.531$, $p = .4522$, $catalogue_{motor-areas} = 0.468$, $p = .6691$, $catalogue_{leftIFG} = 0.5$, $p = .5555$, $catalogue_{leftATL} = 0.562$, $p = .3414$, $catalogue_{leftpSTS} = 0.562$, $p = .3299$; $drawing_{VMPFC} = 0.375$, $p = .9147$, $drawing_{motor-areas} = 0.687$, $p = .0923$, $drawing_{leftIFG} = 0.656$, $p = .0836$, $drawing_{leftATL} = 0.375$, $p = .918$, $drawing_{leftpSTS} = 0.562$, $p = .3108$).

7. Discussion

7.1. Graded concreteness is a powerful dimension for capturing the effect of composition in the brain for verb-noun phrases

The encoding results in Figs. 6 and 7 show that graded concreteness is a powerful lens through which to explore the effect on the brain of verb-noun semantic composition. In particular, the judgments about phrase concreteness consistently provide a good model of the brain response to the stimuli in our study. The simple, single-words concreteness ratings also showed good encoding performance, though the accuracy was lower than with the phrase concreteness model. This confirms previous reports that averaging individual representations, despite being an apparently naïf approach, counts as a solid baseline mechanism when accounting how individual components of meaning are composed in cognition (Anderson et al., 2017; Calvo & Mac Kim, 2013; Dinu et al., 2013; Thornton, Weaverdyck, & Tamir, 2019; Wu et al., 2022).

Overall, these findings support the view that concreteness is an important semantic dimension for these phrases; also, they would seem to provide evidence for concreteness as a continuous, rather than binary variable—a position which has become accepted recently (Borghi et al., 2017; Sakreida et al., 2013).

7.2. ITGPT2 captures fine-grained meaning variation in the brain

Although ITGPT2 did not achieve as high a performance as directly using subjective concreteness ratings, we still found that it had the ability to capture the impact of the stimuli on brain representation well above chance. This suggests that in the long run, contextualized language models may become a potential alternative to rating data, which are difficult and expensive to collect except for small-scale studies (Grand, Blank, Pereira, & Fedorenko, 2022).

The effectiveness of a model like ITGPT2 makes intuitive sense. Modification of the semantic representations through semantic composition is the *raison d'être* of contextualized language models, which represent each word as a complex, nonlinear function of the other words surrounding them. Furthermore, our pattern of results is consistent with recent results in the literature which clearly indicates that contextualized language models provide excellent fit with brain data (Anderson et al., 2021; Bruera & Poesio, 2022; Caucheteux & King, 2022; Goldstein et al., 2022; Jat et al., 2019; Sun, Wang, Zhang, & Zong, 2020; Schrimpf et al., 2021). Note, however, that these previous works only considered either simple concepts or longer, less controlled sentences, while our work focused on an intermediate unit—phrases—and involved much more stringent testing: we used a strictly controlled modulation of concreteness and very specific cases of verb-noun composition. From this point of view, our results provide stronger evidence that contextual language models are able to capture the subtle variations in meaning elicited in the brain, since we focused on an extremely simple case of verb-noun phrase, thus reducing confounds to a minimum.

Our results, however, also point to the current limitations of a model like ITGPT2. Like in the case of concreteness ratings, we compared representations obtained from averaging

individual words—that is, when creating representations composed in a simple, nonlinear method, which has been shown, however, to be quite effective also in NLP (Dinu et al., 2013; Herbelot & Baroni, 2017; Lazaridou, Marelli, & Baroni, 2017)—to the full-fledged, complex, nonlinearly composed phrase representations from contextualized mentions of the phrases. The contrast does indicate an advantage for phrase ITGPT2, the complex model of composition, with higher scores overall when compared to single-words ITGPT2. But the magnitude of such differences is smaller than the one found with concreteness ratings. Furthermore, the model based on full-phrase concreteness is overall better than phrase ITGPT2. The emerging picture, therefore, indicates that extremely controlled fine-grained effects on semantic interpretation triggered by composition are only partially captured by a contextualized language model like ITGPT2. Future developments in the field, which is growing at an extremely fast rate, will show whether such limitations can be overcome by having bigger models, or whether a different approach needs to be taken altogether (Kirstain et al., 2022; Lenci et al., 2022). We believe that, in this respect, our work can also provide relevant insights for NLP—in particular from the point of view of probing whether and how contextualized language models handle composed meaning beyond individual words, such as constructions (Li, Zhu, Thomas, Rudzicz, & Xu, 2022; Madabushi, Romain, Divjak, & Milin, 2020; Veenboer & Bloem, 2023; Weissweiler et al., 2023). With respect to this, the results reported in Appendix D in the Supplementary Materials provide some insights. Converging with recent results showing that creating larger models is not enough to better capture cognitive processing (de Varda & Marelli, 2023; Oh & Schuler, 2023; Oh, Clark, & Schuler, 2022; Shain, Meister, Pimentel, Cotterell, & Levy, 2022), we find that the smallest model (ITGPT2) provides the best encoding performance overall. This seems to confirm that model size may not be the most important factor in order to improve the mapping between contextualized language models and cognitive processing, or at least not for all linguistic phenomena (but see Antonello et al., 2023 for a different opinion). Also, when comparing the patterns of results across layers for single-words and phrase ITGPT2 (Figs. F.1 and F.2 from Appendix F in the Supplementary Materials), it is clear that contextualization is a key factor in determining the differences among the models' representations. For phrase ITGPT2, the complex model of composition, performance increases dramatically as layers progress and as more contextualization takes place (top performance is at layer 23). On the contrary for its simple counterpart, single-words ITGPT2, contextualization is much less important, and the peak of performance is reached much earlier (layer 11).

In other words, our findings provide original evidence with respect to why contextualized language models are able to capture cognitive language processes (for a discussion on the theoretical import of this, see also Günther et al., 2019, and Antonello and & Huth, 2022). These results indicate that one of the key features in modeling linguistic processing in the brain with computational means is modeling complex, nonlinear ways in which constituents need to be modified by the process of composition itself—giving rise to a holistic representation of the phrase whose properties go beyond those of their parts (see further discussion of this view in Section 7.3: Joshi & Schabes, 1997; Kay & Michaelis, 2019; Pustejovsky, 1998).

7.3. Evidence for construction-based semantic composition in the brain

We would argue that our results about learning mappings with concreteness judgments and with computational language models have implications for theoretical accounts of minimal verb-noun composition: specifically, they suggest that meaning resulting from semantic composition in the verb phrases under study is not purely compositional.

The first and main piece of evidence comes from mapping simple and complex composition to and from concreteness models—in the first case, modeling a phrase's semantic representation as a simple average of the individual word's representations (single-words concreteness), and in the second case, assuming that complex, nonlinear semantic composition is carried out by the subjects providing the concreteness ratings for the full phrases (phrase concreteness). The second piece of evidence, which is, however, only confirmatory, as it is less strong, comes from the results with language models. Semantic dimensions are not directly interpretable in such models (Boleda & Erk, 2015), but they do provide the advantage of modeling differently, and more explicitly, the way in which semantic representations compose and are modified by the process of composition.

The results summarized in Fig. 6 show that both with concreteness judgments and language models, modeling semantic composition at the phrase level gives significantly more accurate results than modeling at the word level. This pattern of results emerges very clearly particularly for the cognitive model based on concreteness, indicating that a more holistic model of semantic composition provides a better description of how the brain processes verb-noun composition. Specifically, in the complex (phrase-level) models, the representations of the individual words are modified in complex, nonlinear ways. We hypothesize, therefore, that in verb-noun composition, the interpretation of the constituents is modified by the process of composition itself. In other words, composition does not simply combine the interpretations of the constituents recursively computed bottom up, but may modify or select the semantic representations it acts upon—and therefore, some of the properties of the resulting phrase representation go beyond its parts (Culicover, Jackendoff, & Audring, 2017; Pustejovsky, 1998). This more complex view of composition is consistent with the more recent accounts developed in linguistics, such as the Generative Lexicon (Pustejovsky, 1998), Lexicalized Tree-Adjoining Grammar (Joshi & Schabes, 1997), Construction Grammar (Kay & Michaelis, 2019), or Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1994).

We nevertheless acknowledge that alternative models could have been used for simple composition. This point may be raised especially from the perspective of computational linguistics and NLP, where much research has been dedicated to this subject (see Section 4.2.1). However, it is important to notice that our model of simple semantic composition—based on averaging individual representations—is motivated by two main factors. First of all, it is completely unsupervised, whereas more sophisticated methods used in NLP often make use of supervised learning of composed representations. The main issue about using a supervised approach here would have been choosing the target for training—an a priori "golden" representation of composed semantic representation. This would have introduced circularity in our analyses: it would have required us to assume in advance what a good model of semantic composition in the brain is—which was, however, one of the core questions motivating this

work. Second, among the unsupervised ways of composing semantic representations, averaging provides the main advantages of transparency and effectiveness, which has been proved in both cognitively oriented and NLP-oriented studies (Anderson et al., 2017; Calvo & Mac Kim, 2013; Dinu et al., 2013; Gregori et al., 2020; Herbelot & Baroni, 2017; Lazaridou et al., 2017; Thornton et al., 2019; Wu et al., 2022).

We would like to underline that our results do not rule out that simple composition (so-called "pure" compositionality) can capture *any facet at all* of semantic composition in the brain. In fact, our results indicate that models of simple composition, based on averaging, can actually explain brain activity to a good extent for some cases of composition (see, for instance, Fig. 6). Our main point, however, is different. We believe that our results suggest that, in order to properly explain how fine-grained meaning variation is captured in the brain through semantic composition, a more complex, holistic model of compositionality is needed. In such a model, individual representations are modified by the process of composition itself in complex ways, and the resulting composed meaning goes beyond the sum of its parts. In such a view, the complexity of the mechanisms behind semantic compositionality varies from case to case—and while in some basic instances, pure compositionality can go a long way in explaining cognitive processing, complex compositionality allows to also accommodate more subtle cases, like sense selection, together with the more coarse-grained examples of composition.

7.4. Verb-noun semantic composition in the language network

Brain imaging studies of semantic composition have provided (sometimes contrasting) evidence that different brain regions are involved, including the left inferior frontal gyrus (IFG) (Husband et al., 2011; Schell et al., 2017), the left anterior temporal lobe (ATL) (Pylkkänen, 2020), and the left posterior superior temporal sulcus (pSTS) (Murphy et al., 2022).

In our study, we focus on three linguistically motivated, very specific and distinct cases of semantic composition within the verb-noun semantic composition phrases (sense selection, transparent composition, and light-verb phrases; see Section 3.1 for details). This approach allows us to obtain a detailed picture with respect to verb-noun semantic composition in the brain, revealing which previously reported brain areas support each mode of composition, and consequently, whether each case of composition depends on specific neural processes or not.

As shown in Fig. 8, each ROI contains variable amounts of information related to semantic composition depending on the composition case. This implies that sense selection, transparent composition, and light-verb phrases involve partially different brain processes and resources; such a view, in turn, is compatible with linguistic theories of semantic composition, which propose a similar picture (Jackendoff, 1997; Pustejovsky, 1998). Our results not only confirm the framework provided by theoretical linguists with respect to the differences among these cases, but they also offer ways to reconcile the apparently contradictory accounts of the neural bases of semantic composition. Overall, our results found across the different composition modes are compatible with the presence of a gradient of preferential involvement of different brain regions depending on the characteristics of the specific composition case.

First, the pSTS was found to be the only area strongly involved in all composition cases, with encoding performance being always statistically significant—importantly, it is the only brain area where light-verb phrases could be encoded with statistical significance, and significantly better than all other areas (see Section 6.1). This suggests that it may play a general role in semantic composition processes, with no clear specialization (Matchin & Hickok, 2020; Murphy et al., 2022). Second, with respect to the left ATL, it showed significant performance for the overall and transparent composition cases, and very close to significance for sense selection ($p = .0504$). This is consistent with the view that this area is relevant for modality-independent conceptual representation and combination (Pylkkänen, 2020)—and, we add, holistic modification of the representations through composition (cf. above). Third, regarding the left IFG, encoding reaches statistical significance in all cases except light-verbs, including the polyseme *book*. This involvement in all cases where concreteness is modulated, even in extremely fine-grained cases such as dot-objects, can be explained by proposing that activation patterns in the IFG are differentially modulated by abstract and concrete stimuli, with possible additional effects due to processing of composition (Schell et al., 2017)—an explanation dovetails nicely with a solid body of previous results connecting concreteness and the IFG (e.g., Binder et al., 2005; Della Rosa et al., 2018, reviewed in Bucur & Papagno, 2021).

7.5. Semantic composition outside of the language network

Brain areas outside of the language network have also been associated with semantic composition. We have considered two such cases: the vMPFC, which according to Pylkkänen (2020) should reflect late processing of composition; and a set of motor areas (the precentral gyrus and the SMC) which (Sakreida et al., 2013) found to be activated by modulation of concreteness in verb-noun phrases. The involvement of the two areas would be associated with different functionalities: the vMPFC should be involved in the process of composition itself according to Pylkkänen (2020), whereas motor areas are meant to be involved in the semantic representation of phrase meaning, depending on the level of concreteness (Sakreida et al., 2013)—a hypothesis stemming from the embodied cognition framework (Barsalouet al., 2008), and concurring with recent proposals which describe semantic processing in the brain as an integration of multi-modal information (Binder et al., 2016; Jackson, 2021; Pulvermüller, 2013; Lambon Ralph et al., 2017).

In our analyses on individual brain areas, summarized in Figs. 8 and 9, the vMPFC does not appear to be much involved in semantic composition, since encoding scores are almost never significantly better than chance. This result would appear to be in line with other reports that vMPFC involvement with semantic composition cannot be always replicated (Pylkkänen, 2020). We interpret this as providing further evidence that the role of the vMPFC in semantic composition is not confirmed, and may indeed depend on MEG-specific artifacts or to the task at hand (i.e., whether language production is involved or not)—two explanations proposed in Pylkkänen (2020).

In motor areas, by contrast, we were able to encode brain activity with accuracy above chance not only for all of the composition cases (see Fig. 8; notice that sense selection only

Fig. 9. Encoding brain activity from phrase concreteness for different senses of the same word (representing a dot-object) within the sense selection case across the five regions-of-interest (ROIs). The *Y*-axis represents the accuracy score, averaged across subjects. Bar heights correspond to average values, and average scores for individual subjects are reported as dots. Ceiling values, which quantify the SNR—that is, the level of best possible encoding or decoding of the fMRI data itself, are reported as inverted gray bars, going from 1. to the ceiling accuracy value. The random baseline of 0.5 is indicated by a dotted line. We report the results of statistical tests against the baseline as stars in the lower part of each bar (at a *y* value of slightly above 0.2); one star stands for $p < .05$, two stars for $p < .005$, and three stars for $p < .0005$. Below the plots, we report pairwise statistical comparisons for each possible pair of bars within each section of the plots, using squares whose color reflect the model whose comparison is reported. All *p*-values are FDR-corrected. Results show that brain activity associated with the two senses for *book* can be reliably encoded only in the the left IFG, but performance approaches significance also for the left ATL ($p = .054$). For the word *magazine*, the opposite pattern emerges—encoding performance is statistically significant for all areas except the left IFG. The two senses for *catalogue* and *drawing*, by contrast, cannot be distinguished reliably in any brain area.

approaches statistical significance, with $p = .0668$), but also for the two senses of *magazine* (see Fig. 9). Thus, with respect to the neural bases of semantic processing, our results confirm a graded recruitment of motor areas during language comprehension modulated by fine-grained semantic shifts in concreteness, as already proposed by Sakreida et al. (2013) and fitting with larger pictures of semantics in the brain (Binder et al., 2016; Jackson, 2021; Lambon Ralph et al., 2017).

Notice, however, that our results do not bear directly on the debate on embodied cognition and the role of experiential simulation during language comprehension (Mahon & Caramazza, 2008; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012; Zwaan, 2014). Encoding analyses such as ours only allow us to say that semantic information can explain patterns of activity in motor areas evoked by language. This is compatible with current theories of semantics as a process integrating multi-modal and supra-modal, linguistic features (Lambon Ralph et al.,

2017), but it cannot speak as per *how* and *when* in semantic processing these areas were recruited, something which would instead be needed in order to make a claim in favor of strong theories of embodied cognition (Meteyard et al., 2012).

7.6. Encoding word senses

Finally, the sense selection stimuli give us some—very preliminary—evidence on the extent to which we can retrieve the distinction between word senses for *book*-type dot-objects in the brain. These results are particularly relevant for the debate on the cognitive representation of polysemic lexical items, as it is unclear whether all polysemes should show the same type of cognitive processing, or instead polysemy should work idiosyncratically in each case, possibly interacting with the prototypicality of each noun as a dot-object—for example, *book* would be a more prototypical case of the sense alternation between physical object-informational content than *magazine* (Haber & Poesio, 2021). The results reported in Fig. 7 seem to speak in favor of the latter; notice, however, that they cannot be seen as anything more than as underlined above (see Section 6.2), given the small number of examples.

All polysemous words show different patterns of encoding accuracies, both for ceiling and for the various models, and only some of them can be encoded with accuracy significantly above chance or approaching statistical significance—namely, *book* and *magazine*. This suggests that each polyseme has to be characterized using different models, and possibly different cognitive and neural processes, depending on its specific properties.

Finally, being able to encode the two senses for *book* and *magazine* using GPT-2 indicates that a contextual language model can capture extremely fine-grained information about different senses of words as they are processed in the brain. This represents, to our knowledge, an original finding. This is promising, but further work will be needed to qualify on the one hand, how brain processing changes from one polyseme to another (cf. above), and on the other hand, which types of polysemes are best captured by contextual language models, and why (cf. Haber & Poesio, 2021).

8. Conclusion

During language comprehension, the interpretation of words and phrases varies slightly depending on the context, triggering fine-grained meaning variation, and in some cases, evoking different so-called word senses: for instance, the polysemous noun *book* refers to a physical object when it is preceded by the verb *open*, and instead refers to a piece of information when the verb *copy* is used before it.

In this work, we investigated the neural basis of such fine-grained shifts in lexical semantics modulated by semantic composition in Italian verb-noun phrases. We used a multivariate approach, encoding brain activity using cognitive and computational models. We first compared the performance of four models at encoding brain representations of verb-noun phrases within the language network. Two models were based on cognitive data—concreteness ratings—and two were computational—their representations were extracted from ITGPT2, a contextualized language model for Italian based on the GPT family of models (Radford et al.,

2019). Within each family of models (cognitive and computational), we created representations that captured either *complex* semantic composition, where the individual lexical representations are modified in nonlinear ways during composition, or *simple* composition, where constituents are simply put together through averaging, but not modified.

Then, we also carried out analyses on individual brain areas using the best model (phrase concreteness). We looked at brain areas previously proposed in the literature to be implicated with verb-noun semantic composition, both within the language network (left IFG, left ATL, left pSTS) and outside of it (vMPFC, motor areas).

Our results provide a detailed picture with respect to both theoretical and empirical questions related to semantic composition. Regarding the more theoretical side, comparing different models (Sections 6.1 and 6.2) sheds light on the mechanisms behind semantic composition in the brain. Phrase-based ("complex") models were consistently better than word-based ("simple") models. This seems to suggest that semantic composition in the brain does not simply involve the application of a shallow composition process, but instead modifies in nonlinear ways the representations over which semantic composition operates.

Looking at separate brain areas (Sections 6.3 and 6.4) allowed us, by contrast, to investigate the neural bases of fine-grained semantic shifts triggered by semantic composition. Two brain areas seemed to most consistently contain information regarding semantic composition. The first one was the left pSTS, inside the language network. This confirms its previously proposed role as a generic substrate for combinatory processes. The second one, which is more surprising as it falls outside of the language network, was a set of motor areas (the precentral gyrus and the SMC). This suggests that during semantic composition, modality-specific information (in this case, motor representations of the actions described in the phrase) is recruited. Additionally, the left IFG and the left ATL also seemed to be relevant for semantic composition, although less consistently. Finally, our results add to previous results casting doubts on the role of the vMPFC in semantic composition.

We were also able to encode different senses for some, but not all, polysemous words, both with concreteness models and computational language models—an indication that polysemy can be captured in the brain, but it is a multifaceted and idiosyncratic process.

Acknowledgments

We would like to thank Marco Baroni, Gemma Boleda, Diego Frassinelli, Sabine Schulte im Walde, and the COLT research group for providing insightful feedback on earlier versions of this work.

Ethics statement

All procedures were approved by the ethics committee of the University of Trento, where the data were collected. Participants signed a written consent form before taking part to the experiment, and they received a small monetary compensation at the end of the session.

Notes

1 With a few exceptions, most notably the work by Erk and Padó (2008) and Schütze (1998).
2 The preprocessed fMRI data, the code to replicate the analyses, as well as the figures, evaluations, sentences, and word vectors extracted from language models, can be found at a dedicated repository on the Open Science Foundation website at this link: https://osf.io/sphn4/.
3 The brain masks were downloaded from https://evlab.mit.edu/funcloc/.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging withscikit-learn. *Frontiers in Neuroinformatics*, *14*, 1–10.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., Wang, X., Doko, D., & Raizada, R. D. (2017). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, *27*(9), 4379–4395.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D., Lin, F., & Lalor, E. C. (2019). An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, *39*(45), 8969–8987.

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

Anderson, A. J., Murphy, B., & Poesio, M. (2014). Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*, *26*(3), 658–681.

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *Neuroimage*, *128*, 44–53.

Antonello, R., Vaidya, A., & Huth, A. G. (2023). Scaling laws for language encoding models in fMRI. *arXiv preprint arXiv:2305.11863*.

Antonello, R. J., & Huth, A. (2022). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, 1–16.

Apidianaki, M. (2022). From word types to tokens and back: A survey of approaches to word meaningrepresentation and interpretation. *Computational Linguistics*, *49*(2), 465–523.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, *12*, 5–32.

Baggio, G., Choma, T., Van Lambalgen, M., & Hagoort, P. (2010). Coercion and compositionality. *Journal of Cognitive Neuroscience*, *22*(9), 2131–2140.

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2012). Theprocessing consequences of compositionality. In W. Hinzen, E. Machery & M. Werning (Eds.), *The Oxford handbook of compositionality* (pp. 655–672). Oxford University Press.

Baroni, M. (2013). Composition in distributional semantics. *Language and Linguistics Compass*, *7*(10), 511–522.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, *9*, 241–346.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1183–1193).

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645.

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, *31*(8), 2801–2814.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

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

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*(6), 905–917.

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, *27*(11), 987–989.

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.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213–234.

Boleda, G., & Erk, K. (2015). Distributional semantic features as semantic primitives-or not. In *2015 AAAI Spring Symposium Series*.

Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4758–4781).

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263.

Brennan, J., & Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage*, *60*(2), 1139–1148.

Briem, D., Balliel, B., Rockstroh, B., Butt, M., im Walde, S. S., & Assadollahi, R. (2009). Distinct processing of function verb categories in the human brain. *Brain Research*, *1249*, 173–180.

Bruera, A., & Poesio, M. (2022). Exploring the representations of individual entities in the brain combining EEG and distributional semantics. *Frontiers in Artificial Intelligence*, *5*, 1–25.

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

Brugman, C. (2001). Light verbs and polysemy. *Language Sciences*, *23*(4–5), 551–578.

Bucur, M., & Papagno, C. (2021). An ALE meta-analytical review of the neural correlates of abstract and concrete words. *Scientific Reports*, *11*(1), 1–24.

Butt, M. (2010). The light verb jungle: Still hacking away. *Complex Predicates in Cross-Linguistic Perspective*, 48–78.

Caceres, C. A., Roos, M. J., Rupp, K. M., Milsap, G., Crone, N. E., Wolmetz, M. E., & Ratto, C. R. (2017). Feature selection methods for zero-shot learning of neural activity. *Frontiers in Neuroinformatics*, *11*, 41.

Calvo, R. A., & Mac Kim, S. (2013). Emotions in text: Dimensional and categorical models. *Computational Intelligence*, *29*(3), 527–543.

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, *63*, 743–788.

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning* (pp. 1336–1348). PMLR.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10.

Chersoni, E., Santus, E., Pannitto, L., Lenci, A., Blache, P., & Huang, C.-R. (2019). A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, *25*(4), 483–502.

Chyzhyk, D., Varoquaux, G., Milham, M., & Thirion, B. (2022). How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*, *11*, 1–15.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory* (pp. 493–522).

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLoS One*, *12*(11), e0184923.

Cruse, D. A. (1992). Monosemy vs. polysemy. *Linguistics*, *30*(3), 577–599.

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, *128*(3), 615–627.

Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword constructions in the grammar. *Topics in Cognitive Science*, *9*(3), 552–568.

De Almeida, R. G., Riven, L., Manouilidou, C., Lungu, O., Dwivedi, V. D., Jarema, G., & Gillon, B. (2016). The neuronal correlates of indeterminate sentence comprehension: An fMRI study. *Frontiers in Human Neuroscience*, *10*, 614.

de Varda, A., & Marelli, M. (2023). Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 139–149). Toronto, Canada: Association for Computational Linguistics.

de Vries, W., & Nissim, M. (2021). As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 836–846).

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., Zevin, J. D., Immordino-Yang, M. H., Gordon, A. S., Damasio, A., & Kaplan, J. T. (2017). Decoding the neural representation of story meanings across languages. *Human Brain Mapping*, *38*(12), 6096–6106.

Delgado, M. R., Beer, J. S., Fellows, L. K., Huettel, S. A., Platt, M. L., Quirk, G. J., & Schiller, D. (2016). Viewpoints: Dialogues on the functional role of the ventromedial prefrontal cortex. *Nature Neuroscience*, *19*(12), 1545–1552.

Della Rosa, P. A., Catricalà, E., Canini, M., Vigliocco, G., & Cappa, S. F. (2018). The left inferior frontal gyrus: A neural crossroads between abstract and concrete knowledge. *Neuroimage*, *175*, 449–459.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert M. S., & Killiany R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, *118*(2), 115–128.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186).

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, *13*(4), e1005508.

Dinu, G., Pham N. T., & Baroni, M.(2013). General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality* (pp. 50–58). Sofia, Bulgaria. Association for Computational Linguistics.

Djokic, V. G., Maillard, J., Bulat, L., & Shutova, E. (2020). Decoding brain activity associated with literal and metaphoric sentence comprehension using distributional semantic models. *Transactions of the Association for Computational Linguistics*, *8*, 231–246.

Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1325–1335).

Emmanuele, C., Enrico, S., Chu-Ren, H., & Lenci, A. (2021). Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, *47*(3), 663–698.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Erk, K., & Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 897–906).

Falkum, I. L., & Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua*, *157*, 1–16.

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology*, *61*(6), 825–850.

Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781.

Fodor, J., & Lepore, E. (2000). The emptiness of the lexicon: Critical reflections on J. Pustejovsky's the generative lexicon. In *Meaning and the lexicon*. New York: Crowell.

Frankland, S. M., & Greene, J. D. (2020). Two ways to build a thought: Distinct forms of compositional semantic representation across brain regions. *Cerebral Cortex*, *30*(6), 3838–3855.

Frassinelli, D., & Im Walde, S. S. (2019). Distributional interaction of concreteness and abstractness in verb–noun subcategorisation. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers* (pp. 38–43).

Frassinelli, D., Naumann, D., Utt, J., & m Walde, S. S. (2017). Contextual characteristics of concrete and abstract words. In *IWCS 2017-12th International Conference on Computational Semantics-Short papers*.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*, 25–50.

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254.

Frisson, S. (2015). About bound and scary books: The processing of book polysemies. *Lingua*, *157*, 17–35.

Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: Distinct ERP effects of intensionality and denotation. *Language, Cognition and Neuroscience*, *35*(10), 1295–1313.

Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N., & Mitchell, T. M. (2019). The lexical semantics of adjective–noun phrases in the human brain. *Human Brain Mapping*, *40*(15), 4457–4469.

Ghio, M., Vaghi, M. M. S., & Tettamanti, M. (2013). Fine-grained semantic categorization across the abstract and concrete domains. *PLoS One*, *8*(6), e67090.

Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *Quarterly Journal of Experimental Psychology*, *61*(6), 905–919.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen A., Gazula H., Choe G., Rao A., Kim C., Casto C., Fanda L., Doyle W, Friedman D., Dugan P., Melloni L., Reichart R., Devore S., Flinker A., Hasenfratz L., Levy O., Hassidim A., Brenner M., Matias Y., Norman K. A., Devinsky O., & Hasson U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Gorman, K., & Bedrick, S. (2019). We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2786–2791).

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987.

Grefenstette, E., Dinu, G., Zhang, Y.-Z., Sadrzadeh, M., & Baroni, M. (2013). Multi-step regression learning for compositional distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers* (pp. 131–142).

Gregori, L., Montefinese, M., Radicioni, D. P., Ravelli, A. A., Varvara, R., et al. (2020). Concretext@ evalita2020: The concreteness in context task. In *EVALITA*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Haber, J., & Poesio, M. (2021). Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2663–2676).

Haber, J., & Poesio, M. (2023). Polysemy - Evidence from linguistics, behavioural science and contextualised language models. *Computational Linguistics*.

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, *8*(1), 427–446.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, *37*(1), 435–456.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*(2), 257–270.

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, *180*, 4–18.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus–It's my area: Or is it? *Journal of Cognitive Neuroscience*, *20*(12), 2125–2136.

Herbelot, A., & Baroni, M. (2017). High-risk learning: Acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hill, F., & Korhonen, A. (2014). Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 725–731).

Hill, F., Korhonen, A., & Bentz, C. (2014). A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, *38*(1), 162–177.

Honari-Jahromi, M., Chouinard, B., Blanco-Elorrieta, E., Pylkkänen, L., & Fyshe, A. (2021). Neural representation of words within phrases: Temporal evolution of color-adjectives and object-nouns during simple composition. *PLoS One*, *16*(3), e0242754.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, *119*, 456–467.

Husband, E. M., Kelly, L. A., & Zhu, D. C. (2011). Using complement coercion to understand the neural basis of semantic composition: Evidence from an fMRI study. *Journal of Cognitive Neuroscience*, *23*(11), 3254–3266.

Izsak, P., Berchansky, M., & Levy, O. (2021). How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10644–10652).

Jackendoff, R. (1997). *The architecture of the language faculty*. Number 28. MIT Press.

Jackson, R. L. (2021). The neural correlates of semantic control revisited. *Neuroimage*, *224*, 117444.

Jat, S., Tang, H., Talukdar, P., & Mitchell, T. (2019). Relating simple sentence representations in deep neural networks and the brain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5137–5154).

Joshi, A. K., & Schabes, Y. (1997). Tree-adjoining grammars. In G. Rozenberg & A. Salomaa (Eds.), *Handbook of formal languages* (pp. 69–123). Springer.

Kaiser, D., Jacobs, A. M., & Cichy, R. M. (2022). Modelling brain representations of abstract concepts. *PLoS Computational Biology*, *18*(2), e1009837.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Katsika, A., Braze, D., Deo, A., & Piñango, M. M. (2012). Complement coercion: Distinguishing between type-shifting and pragmatic inferencing. *Mental Lexicon*, *7*(1), 58–76.

Kay, K. N. (2018). Principles for models of neural information processing. *Neuroimage*, *180*, 101–109.

Kay, P., & Michaelis, L. A. (2019). *Constructional meaning and compositionality* (pp. 293–324). De Gruyter Mouton.

Kirstain, Y., Lewis, P., Riedel, S., & Levy, O. (2022). A few more examples may be worth billions of parameters. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 1017–1029).

Klepousniotou, E., Gracco, V. L., & Pike, G. B. (2014). Pathways to lexical ambiguity: fMRI evidence for bilateral fronto-parietal involvement in language processing. *Brain and Language*, *131*, 56–64.

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*, *123*(1), 11–21.

Kochari, A. R., Lewis, A. G., Schoffelen, J.-M., & Schriefers, H. (2021). Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: An MEG study. *Neuropsychologia*, *155*, 107754.

Kragel, P. A., Koban, L., Barrett, L. F., & Wager, T. D. (2018). Representation, pattern information, and brain signatures: From neurons to neuroimaging. *Neuron*, *99*(2), 257–273.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis–Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *4*, 1–28.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80.

Kuperberg, G. R., Choi, A., Cohn, N., Paczynski, M., & Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience*, *22*(12), 2685–2701.

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning* (pp. 2873–2882). PMLR.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lauwers, P., & Willems, D. (2011). Coercion: Definition and challenges, current approaches, and new trends. *Linguistics*, *49*(6), 1219–1235.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*, 677–705.

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage*, *56*(2), 387–399.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 58–66).

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, *56*, 1269–1313.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302–308).

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Li, B., Zhu, Z., Thomas, G., Rudzicz, F., & Xu, Y. (2022). Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7410–7423).

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru R., Shleifer S., Koura P. S., Chaudhary V., O'Horo B., Wang J., Zettlemoyer L., Kozareva Z., Diab M., Stoyanov V., & Li X. (2021). Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Lukic, S., Meltzer-Asscher, A., Higgins, J., Parrish, T. B., & Thompson, C. K. (2019). Neurocognitive correlates of category ambiguous verb processing: The single versus dual lexical entry hypotheses. *Brain and Language*, *194*, 65–76.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Lyons, J. (1977). *Semantics: Volume 2*. Cambridge University Press.

MacGregor, L. J., Bouwsema, J., & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia*, *68*, 126–138.

Madabushi, H. T., Romain, L., Divjak, D., & Milin, P. (2020). CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4020–4032).

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, *102*(1–3), 59–70.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Matchin, W., & Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, *30*(3), 1481–1498.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 1–9.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *57*(1), 145–175.

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.

Mkrtychian, N., Blagovechtchenski, E., Kurmakaeva, D., Gnedykh, D., Kostromina, S., & Shtyrov, Y. (2019). Concrete vs. abstract semantics: From mental representations to functional brain mapping. *Frontiers in Human Neuroscience*, *13*, 267.

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., Kean, H., Qian, P., & Fedorenko, E. (2020). Composition is the core driver of the language-selective network. *Neurobiology of Language*, *1*(1), 104–134.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. M. E. Moravcsik & P. Suppes (Eds.), *Approaches to natural language* (pp. 221–242). Springer.

Montefinese, M. (2019). Semantic representation of abstract and concrete words: A minireview of neural evidence. *Journal of Neurophysiology*, *121*(5), 1585–1587.
Murphy, B., Baroni, M., & Poesio, M. (2009). EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of EMNLP* (pp. 619–627). Singapore.
Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., & Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, *117*(1), 12–22.
Murphy, B., Wehbe, L., & Fyshe, A. (2018). Decoding language from the brain. In T. Poibeau & A. Villavicencio (Eds.), *Language, cognition, and computational models* (pp. 53–80).
Murphy, E., Woolnough, O., Rollo, P. S., Roccaforte, Z. J., Segaert, K., Hagoort, P., & Tandon, N. (2022). Minimal phrase composition revealed by intracranial recordings. *Journal of Neuroscience*, *42*(15), 3216–3227.
Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences*, *19*(10), 551–554.
Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400–410.
Naumann, D., Frassinelli, D., & Schulte im Walde, S. (2018). Quantitative semantic variation in the contexts of concrete and abstract words. In *Seventh Joint Conference on Lexical and Computational Semantics (SEM 2018)* (pp. 76–85).
Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (2003). *Polysemy: Flexible patterns of meaning in mind and language*. De Gruyter.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), e1003553.
Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, 777963.
Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.
Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241.
Partee, B. H. (2008). *Compositionality in formal semantics: Selected papers*. John Wiley & Sons.
Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).
Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier.
Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 1–13.
Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, *45*(1), S199–S209.
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). 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).
Pinkal, M. (1995). *Logic and lexicon. The semantics of the indefinite*. Dordrecht: Kluwer Academic Publishers.
Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago, CA: University of Chicago Press.
Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, *35*(7), 3276–3284.
Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, *17*(9), 458–470.
Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, *17*(4), 409–441.
Pustejovsky, J. (1998). *The generative lexicon*. MIT Press.
Pustejovsky, J. (2011). Coercion in a general theory of argument selection. *Linguistics*, *49*(6), 1401–1431.

Pustejovsky, J., Rumshisky, A., Plotnick, A., Ježek, E., Batiukova, O., & Quochi, V. (2010). Semeval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 27–32).

Pylkkänen, L. (2008). Mismatching meanings in brain and behavior. *Language and Linguistics Compass*, *2*(4), 712–738.

Pylkkänen, L. (2020). Neural basis of basic composition: What we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B*, *375*(1791), 20190299.

Pylkkänen, L., Llinás, R., & McElree, B. (2004). Distinct effects of semantic plausibility and semantic composition in MEG. In *Biomag 2004: Proceedings of the 14th International Conference on Biomagnetism*. Boston, USA: Citeseer.

Pylkkänen, L., Llinás, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, *18*(1), 97–109.

Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 539–579). Elsevier.

Pylkkänen, L., & McElree, B. (2007). An MEG study of silent meaning. *Journal of Cognitive Neuroscience*, *19*(11), 1905–1921.

Pylkkänen, L., Oliveri, B., & Smart, A. J. (2009). Semantics vs. world knowledge in prefrontal cortex. *Language and Cognitive Processes*, *24*(9), 1313–1334.

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

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

Rybář, M., & Daly, I. (2022). Neural decoding of semantic concepts: A systematic literature review. *Journal of Neural Engineering*, *19*, 021002.

Sakreida, K., Scorolli, C., Menz, M. M., Heim, S., Borghi, A. M., & Binkofski, F. (2013). Are abstract action words embodied? An fMRI investigation at the interface between language and motor cognition. *Frontiers in Human Neuroscience*, *125*, 1–13.

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex*, *96*, 105–120.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022). Large-scale evidence for logarithmic effects of word predictability on reading time.

Snell, J., & Grainger, J. (2017). The sentence superiority effect revisited. *Cognition*, *168*, 217–221.

Stkapor, K. (2017). Evaluating and comparing classifiers: Review, some recommendations and limitations. In *International Conference on Computer Recognition Systems* (pp. 12–21). Springer.

Suárez, L. E., Markello, R. D., Betzel, R. F., & Misic, B. (2020). Linking structure and function in macroscale brain networks. *Trends in Cognitive Sciences*, *24*(4), 302–315.

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *Neuroimage*, *62*(1), 451–463.

Sun, J., Wang, S., Zhang, J., & Zong, C. (2020). Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 589–603.

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The brain represents people as the mental states they habitually experience. *Nature Communications*, *10*(1), 2291.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Lrec*, volume 2012 (pp. 2214–2218). Citeseer.

Troche, J., Crutch, S. J., & Reilly, J. (2017). Defining a conceptual topography of word concreteness: Clustering properties of emotion, sensation, and magnitude among 750 English words. *Frontiers in Psychology*, *8*, 1787.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*(6), e12844.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 1–11.

Veenboer, T., & Bloem, J. (2023). Using collostructional analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 12937–12951).

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K. A., & Arora, S. (2018). Mapping between fMRI responses to movies and their natural language annotations. *Neuroimage*, *180*, 223–231.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7222–7240).

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

Weissweiler, L., He, T., Otani, N., Mortensen, D. R., Levin, L., & Schütze, H. (2023). Construction grammar provides unique insight into neural language models. In *Proceedings of the 1st International Workshop on Construction Grammars and NLP (CxGs+ NLP, GURT/SyntaxFest 2023)* (pp. 85–95).

Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and Language*, *141*, 124–134.

Wittenberg, E., & Levy, R. (2017). If you want a quick kiss, make it count: How choice of syntactic construction affects event construal. *Journal of Memory and Language*, *94*, 254–271.

Wittenberg, E., Paczynski, M., Wiese, H., Jackendoff, R., & Kuperberg, G. (2014). The difference between "giving a rose" and "giving a kiss": Sustained neural activity to the light verb construction. *Journal of Memory and Language*, *73*, 31–42.

Wu, M.-H., Anderson, A. J., Jacobs, R. A., & Raizada, R. D. (2022). Analogy-related information can be accessed by simple addition and subtraction of fMRI activation patterns, without participants performing any analogy task. *Neurobiology of Language*, *3*(1), 1–17.

Zaccarella, E., Meyer, L., Makuuchi, M., & Friederici, A. D. (2017). Building by syntax: The neural basis of minimal linguistic structures. *Cerebral Cortex*, *27*(1), 411–421.

Zarcone, A., McRae, K., Lenci, A., & Padó, S. (2017). Complement coercion: The joint effects of type and typicality. *Frontiers in Psychology*, *8*, 1987.

Zhang, L., & Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study. *Neuroimage*, *111*, 228–240.

Zhang, L., & Pylkkänen, L. (2018). Semantic composition of sentences word by word: MEG evidence for shared processing of conceptual and logical elements. *Neuropsychologia*, *119*, 392–404.

Zhang, Y., Warstadt, A., Li, X., & Bowman, S. (2021). When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1112–1125).

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation*, *44*, 35–62.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. A.1: Encoding scores for the various composition cases using a whole-brain analysis.

Fig. A.2: Encoding scores for polysemy cases using a whole-brain analysis.

Fig. A.3: Features used for encoding (across subjects) using the whole-brain approach.

Fig. B.1: Time-resolved encoding scores for phrase concreteness.

Fig. B.2: Time-resolved encoding scores for phrase ITGPT2.

Fig. B.3: Time-resolved encoding scores for single-words concreteness.

Fig. B.4: Time-resolved encoding scores for single-words ITGPT2.

Fig. C.1: Encoding scores for the various composition cases using ITGPT2 with different combinations of corpora as sources for the sentences to be used for representation pooling.

Fig. D.1: Encoding scores for the various composition cases using progressively bigger versions of XGLM.

Fig. E.1: Encoding scores for the various composition cases using fasttext, a state-of-the-art static language model.

Fig. F.1: Layer-by layer encoding scores for the various composition cases using phrase ITGPT2.

Fig. F.2: Layer-by layer encoding scores for the various composition cases using single-words ITGPT2.