Decoding Word Embeddings with Brain-Based Semantic Features

Emmanuele Chersoni
The Hong Kong Polytechnic University
Department of Chinese and
Bilingual Studies
emmanuele.chersoni@polyu.edu.hk

Enrico Santus
MIT Computer Science and Artificial
Intelligence Laboratory
esantus@mit.edu

Chu-Ren Huang
The Hong Kong Polytechnic University
Department of Chinese and
Bilingual Studies
churen.huang@polyu.edu.hk

Alessandro Lenci
University of Pisa
Department of Philology, Literature
and Linguistics
alessandro.lenci@unipi.it

Word embeddings are vectorial semantic representations built with either counting or predicting techniques aimed at capturing shades of meaning from word co-occurrences. Since their introduction, these representations have been criticized for lacking interpretable dimensions. This property of word embeddings limits our understanding of the semantic features they actually encode. Moreover, it contributes to the "black box" nature of the tasks in which they are used, since the reasons for word embedding performance often remain opaque to humans. In this contribution, we explore the semantic properties encoded in word embeddings by mapping them onto interpretable vectors, consisting of explicit and neurobiologically motivated semantic features (Binder et al. 2016). Our exploration takes into account different types of embeddings, including factorized count vectors and predict models (Skip-Gram, GloVe, etc.), as well as the most recent contextualized representations (i.e., ELMo and BERT).

Submission received: 12 February 2020; revised version received: 23 June 2021; accepted for publication: 26 June 2021.

https://doi.org/10.1162/COLI_a_00412

In our analysis, we first evaluate the quality of the mapping in a retrieval task, then we shed light on the semantic features that are better encoded in each embedding type. A large number of probing tasks is finally set to assess how the original and the mapped embeddings perform in discriminating semantic categories. For each probing task, we identify the most relevant semantic features and we show that there is a correlation between the embedding performance and how they encode those features. This study sets itself as a step forward in understanding which aspects of meaning are captured by vector spaces, by proposing a new and simple method to carve human-interpretable semantic representations from distributional vectors.

1. Introduction

One of the most influential and longstanding approaches to semantic representation assumes that the conceptual content of lexical items is decomposable into **semantic features** that identify meaning components, hence the name of **featural**, **decompositional**, or **componential** theories of meaning (Vigliocco and Vinson 2007). In linguistics, features are typically represented by symbols (e.g., HUMAN, PATH, CAUSE) standing for basic or primitive semantic dimensions (Jackendoff 1990; Wierzbicka 1996; Murphy 2010; Pustejovsky and Batiukova 2019). These "building blocks" of meaning are selected a priori and structured into categorical representations defined by the presence or absence of symbolic features, as in this semantic analysis of *enter*:

(1) *enter* [+MOVE, +PATH, -CAUSE, . . .]

Besides the issue of establishing the criteria to define the repertoire of alleged semantic primitives, discrete symbolic structures strive to cope with the gradient nature of lexical meaning and cannot capture the varying degrees of feature prototypicality in concepts (Murphy 2002). Second, the basic semantic features are normally too coarse-grained to provide a full characterization of conceptual content (e.g., accounting for the dimensions that distinguish *painter* from *violinist*). In cognitive psychology, instead of using categorical representations formed by hand-selected components, it is customary to represent concepts with verbal properties generated by native speakers to describe a word meaning and collected in feature norms (e.g., McRae et al. 2005; Vinson and Vigliocco 2008; Devereux et al. 2014). Each feature is associated with a weight corresponding to the number of subjects that listed it for a given concept and is used to estimate its salience in that concept. The following is a representation of *car* using a subset of its feature distribution from the norms in McRae et al. (2005):

(2)

	a_vehicle	*has_4_wheels*	*is_fast*	*is_expensive*
car	9	18	9	11

The main advantage of featural representations is that they are *human-interpretable* and *explainable*: Features explicitly label the dimensions of word meanings and provide explanatory factors of their semantic behavior (e.g., the similarity between *violinist* and *athlete* can be explained by assuming that they both share the feature HUMAN). Conversely, featural semantic representations raise several methodological concerns, as they are either based on intuition and therefore highly subjective, or must be carried out with a complex and time-consuming process of elicitation from human subjects, which is hardly scalable to cover large areas of the lexicon. In fact, existing feature norms only include some hundreds of lexical items, typically limited to concrete nouns.

Semantic features have been widely used in computational linguistics and artificial intelligence (AI), but their limits have eventually contributed to the success of a completely different approach to semantic representation. This is based on data-driven, low-dimensional, dense distributional vectors called **word embeddings**, which represent lexical meaning as a function of the statistical distribution of words in texts. Word embeddings are built by Distributional Semantic Models (DSMs) (Turney and Pantel 2010; Lenci 2018) using various types of methods, ranging from the factorization of co-occurrence matrices with Singular Value Decomposition (SVD) to neural language models. Traditional DSMs have represented the content of lexical types through a single vector that "summarizes" their whole distributional history, disregarding that word tokens may have different meanings in different contexts. Things have recently changed with the introduction of deep neural architectures for language modeling such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019), whose word representations have helped achieve state-of-the-art results in a wide variety of supervised natural language processing (NLP) tasks. These embeddings are "contextualized," in the sense that the model computes a different vector for each token of the same word, depending on the sentence in which it occurs. The popularity of word embeddings, both contextualized and not, is due to the fact that they allow for the fast construction of continuous semantic representations that can be pretrained on large natural language corpora. The vector-based encoding of meaning is easily *machine-interpretable*, as embeddings can be directly fed into complex neural architectures and indeed boost performance in several NLP tasks and applications.

Although word embeddings play an important role in the success of deep learning models and do capture some aspects of lexical meaning, it is hard to understand their actual semantic content. In fact, one notorious problem of embeddings is their lack of *human-interpretability*: Information is distributed across vector dimensions that cannot be labeled with specific semantic values. In neural word embeddings, the vector dimensions are learned as network parameters, instead of being derived from explicit co-occurrence counts between target words and linguistic contexts, making their interpretation even more challenging. Scholars have argued that DSMs provide a *holistic* representation of meaning, as the content of each word can exclusively be read off from its position relative to other elements in the semantic space, while the coordinates of such space are themselves arbitrary and without any intrinsic semantic value (Landauer et al. 2007; Vigliocco and Vinson 2007; Sahlgren 2008). This makes embeddings **"black box" representations** that can be understood only by observing their behavior in some external task, but whose internal content defies direct inspection. A recent and widely used tool to investigate the linguistic properties captured by embeddings are the so-called probing tasks (Ettinger, Elgohary, and Resnik 2016; Adi et al. 2017; Conneau et al. 2018; Kann et al. 2019). A **probing task** is a classification problem that targets a specific linguistic aspect (word order, animacy, etc.). The name refers to the fact that the classifier is used to "probe" embeddings for a particular piece of linguistic information. The successful performance of an embedding model to address this task is then used to infer that the vectors encode that information. However, as it was recently pointed out by Shwartz and Dagan (2019), probing tasks are also a form of "black box" testing, since they just provide indirect evidence about the embedding content.

The emergence of the *interpretability problem* in AI and NLP has motivated the necessity of understanding which shades of semantics are actually encoded by word embeddings, and has therefore refueled the debate about the relationship between distributional representations and semantic features (Boleda and Erk 2015). "Opening the black box" of deep learning methods has become an imperative in computational

linguistics (Linzen, Chrupała, and Alishahi 2018; Linzen et al. 2019). Such research effort aims at analyzing the specific information encoded by vector representations that may help explain their behavior in downstream tasks and applications.

In this article, we contribute to this goal by showing that featural semantic representations can be used to interpret the content of word embeddings. In particular, we argue that decoding semantic information from distributional vectors is strikingly similar to the problem faced by neuroscience of how to "read off meaning" from distributed brain activity patterns. **Neurosemantic decoding** is a research line that develops computational methods to identify the mental state represented by brain activity recorded with neuroimaging techniques such as functional magnetic resonance imaging (fMRI) (e.g., recognizing that a given activation pattern produced by a stimulus picture or word corresponds to an apple). A common approach to address such tasks is to learn a mapping between featural concept representations and a vector containing the corresponding fMRI recorded brain activity (Naselaris et al. 2011; Poldrack 2011). These computational models are able to predict the concept corresponding to a certain brain activation and contribute to shedding light on the neural representation of semantic features.

In neurosemantic decoding, human-interpretable semantic vectors are used to decode the content of vectors of "brain-interpretable" signals activated by a certain stimulus (cf. Section 2.2). In a similar way, we aim at decoding the semantic content of word embeddings by learning a mapping onto vectors of human-interpretable features. To this end, we use the semantic features introduced by Binder et al. (2016), who proposed a set of cognitively motivated semantic primitives (henceforth, **Binder features**) derived from a wide variety of modalities of neural information processing (hence their definition as *brain-based*), and provided human ratings about the relevance of each feature for a set of English words (henceforth, **Binder data set**). We use these ratings to represent the words with continuous vectors of semantic features and to learn a map from word embeddings dimensions to Binder features. Such mapping provides a human-interpretable correlate of word embeddings that we use to address these issues:

1. identifying which semantic features are best encoded in word embeddings;
2. explaining the performance of embeddings in semantic probing tasks.

The idea of mapping word embeddings onto semantic features is not by itself new (Făgărășan, Vecchi, and Clark 2015; Utsumi 2020), but to the best of our knowledge the present contribution is the first one to use mapped featural representations to interpret the semantic information learned by probing classifiers and to explain the embedding behavior in such tasks. Therefore, we establish a bridge between the research on semantic features and the challenge of enhancing the interpretability of distributed representations, by showing that featural semantic representations can work as an important key to open the black boxes of word embeddings and of the tasks in which they are used. As an additional element of novelty, we also apply the neural decoding methodology to the recently introduced contextualized embeddings, to evaluate whether and how they differ from static ones in encoding semantic information. It is important to point out that we do not argue that Binder feature vectors should replace distributional representations. The main claim of this article is rather that continuous vectors of human-interpretable semantic features, such as Binder's, are an important tool to investigate what aspects of meaning distributional embeddings actually encode, and they can be used to lay a bridge between symbolic and distributed semantic representations.

This article is organized as follows. Section 2 introduces the main typologies of DSMs and reviews the related work on vector decoding. In Section 3, we describe the Binder features, we present the method used to map word embeddings onto Binder feature vectors, and we evaluate the mapping accuracy. In Section 4, we investigate which Binder features are best encoded by each type of embedding. In Section 5 we set up a series of probing tasks to verify how the original and mapped embeddings encode semantic categories, such as animate/inanimate or positive/negative sentiment. Some probing tasks focus on static embeddings, whereas others target the token vectors produced by contextualized embeddings. The aim of the analysis is to identify the most important semantic features for a given task and to investigate whether there is a correlation between the system performance and how those features are encoded by the embeddings.

2. Related Work

2.1 From Static Distributional Models to Contextualized Embeddings

We use the term **word embedding** to refer to any kind of dense, low-dimensional distributional vector. In the early days of Distributional Semantics, embeddings were built by applying dimensionality reduction techniques derived from linear algebra, such as SVD, to matrices keeping track of the co-occurrence information about the target terms and some predefined set of linguistic contexts. Parameter tuning was mostly carried out empirically, as it was driven by the model performance on specific tasks. This family of DSMs is referred to as **count models** (Baroni, Dinu, and Kruszewski 2014).

The construction of distributional representations started to be conceived mainly as the byproduct of a supervised language modeling task after the introduction of the Word2Vec package (Mikolov et al. 2013). Low-dimensional distributional word vectors are created by neural network algorithms by learning to optimally predict the contexts of a target word, hence their name, **predict models**. "Neural" embeddings have become an essential component for several NLP applications, also thanks to the availability of many efficient and easy-to-use tools (Mikolov et al. 2013; Bojanowski et al. 2017) that allow researchers to quickly obtain well-performing word representations. Indeed, an important finding of a first comparative evaluation between count and predict models was that the latter achieve far superior performances in a wide variety of tasks (Baroni, Dinu, and Kruszewski 2014). Although the result was claimed to be due to the suboptimal choice of "vanilla" hyperparameters for the count models (Levy, Goldberg, and Dagan 2015), it was still proof that predict models could be very efficient even without any parameter tuning. Subsequent studies adopting cognitively motivated benchmarks (e.g., based on priming, eye-tracking, or electroencephalogram data) have also showed that word embeddings exhibit strong correlation with human performance in psycholinguistic and neurolinguistic tasks (Søgaard 2016; Mandera, Keuleers, and Brysbaert 2017; Bakarov 2018; Schwartz and Mitchell 2019; Hollenstein et al. 2019). Finally, and significantly, Carota et al. (2017) found that the semantic similarity computed via distributional models between action-related words correlates with the fMRI response patterns of the brain regions that are involved in the processing of this category of lexical items.

Another novelty recently came out from the research on deep neural networks for language modeling. For both count and predict models, a common and longstanding assumption was the building of a single, stable representation for each word type in the corpus. In the latest generation of embeddings, instead, each occurrence of a word in a

specific sentence context gets a unique representation (Peters et al. 2018). Such models typically rely on an encoder (i.e., a LSTM or a Transformer) trained on large amounts of textual data, and the word vectors are learned as a function of the internal states of the encoder, such that a word in different sentence contexts determines different activation states and is represented by a distinct vector (McCann et al. 2017; Peters et al. 2018; Howard and Ruder 2018; Devlin et al. 2019; Yang et al. 2019). Thus, the embeddings produced by these new frameworks are said to be **contextualized**, as opposed to the **static** vectors produced by the earlier frameworks, and they aim at modeling the specific sense assumed by the word in context (Wiedemann et al. 2019). Interestingly, the distinction between traditional and contextualized embeddings has been recently discussed by drawing a parallel between the *prototype* and *exemplar* models of categorization in cognitive psychology (Sikos and Padó 2019).

Two very popular models for obtaining contextualized word embeddings are ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019). ELMo is based on a two-layer LSTM trained as the concatenation of a forward and a backward language model, BERT on a stack of Transformer layers (Vaswani et al. 2017) trained jointly in a masked language model and a next sentence prediction task. The semantic interpretation of the dimensions of contextualized embeddings is still an open question. The classical approach to analyze the syntactic and semantic information encoded by these representations is to test them in some probing tasks (Tenney et al. 2019; Liu et al. 2019; Hewitt and Manning 2019; Kim et al. 2019; Kann et al. 2019; Yaghoobzadeh et al. 2019; Jawahar et al. 2019). In this contribution we adopt a different approach to the problem, mainly inspired by the literature on neurosemantic decoding.

2.2 Interpreting Vector Representations

Like word embeddings, the brain encodes information in distributed activity patterns that defy direct interpretation. The general goal of **neurosemantic decoding** is to develop computational methods to infer the content of brain activities associated with a certain word or phrase (e.g., to recognize that a pattern of brain activations corresponds to the meaning of the stimulus word *dog*, instead of *car*). One of the most common approaches to neural decoding consists in learning to map vectors of fMRI signals onto vectors of semantic dimensions. If the mapping is successful, we can infer that these dimensions are encoded in the brain. Mitchell et al. (2008) pioneered such a method by training a linear regression model from a set of words with their fMRIs. The trained model was then asked to predict the activations for unseen words. Approaches differ for the type of semantic representation adopted to model brain data. Mitchell et al. (2008) used a vector of features corresponding to textual co-occurrences with 25 verbs capturing basic semantic dimensions (e.g., *hear*, *eat*). Chang, Mitchell, and Just (2011) instead represented words with vectors of verbal properties derived from feature norms, and Anderson et al. (2016) with vectors of Binder features (cf. Section 3.2).

After the popularization of DSMs, the use of word embeddings for neurosemantic decoding has become widespread. Actually, the decoding task itself has turned into an important benchmark for DSMs, since it is claimed to represent a more robust alternative to the traditional use of behavioral data sets (Murphy, Talukdar, and Mitchell 2012). Some of these studies used fMRI data to learn a mapping from the classical count-based distributional models (Devereux, Kelly, and Korhonen 2010; Murphy, Talukdar, and Mitchell 2012), from both count and prediction vectors (Bulat, Clark, and Shutova 2017b; Abnar et al. 2018), from contextualized vectors (Beinborn, Abnar, and Choenni 2019), or from topic models (Pereira, Detre, and Botvinick 2011; Pereira, Botvinick, and

Detre 2013). This methodology has recently been extended beyond words to represent the meanings of entire sentences (Anderson et al. 2016; Pereira et al. 2018; Sun et al. 2019), even in the presence of complex compositionality phenomena such as negation (Djokic et al. 2019), or to predict the neural responses to complex visual stimuli (Güçlü and van Gerven 2015). Athanasiou, Iosif, and Potamianos (2018) showed that neural activation semantic models built out of these mappings can also be used to successfully carry out NLP tasks such as similarity estimation, concept categorization, and textual entailment.

Despite the analogy, it is important to underline a crucial difference between our work and neurosemantic decoding. In the latter, word embeddings are used as proxies for semantic representations to decode brain patterns that are not directly human-interpretable. Our aim is instead to decode the content of word embeddings themselves. We actually believe this enterprise to be also relevant for (and to a certain extent a precondition to) the task of decoding brain states. In fact, if we want to use embeddings for neural decoding, it is essential to have a better understanding of the semantic content hidden in distributional representations. Otherwise, the risk is to run into the classical fallacy of *obscurum per obscurius*, in which one tries to explain something unknown (brain activations), with something that is even less known (word embeddings).

Another related line of work makes use of property norms for grounding distributional models in perceptual data, and to map them onto interpretable representations (Făgărășan, Vecchi, and Clark 2015; Bulat, Kiela, and Clark 2016; Derby, Miller, and Devereux 2019), an approach that has been proven useful, among other things, also for the detection of cross-domain mappings in metaphors (Bulat, Clark, and Shutova 2017a). Similarly, other studies focused on conceptual categorization proposed to learn mappings from distributional vectors to spaces of higher-order concepts (Şenel et al. 2018; Schwarzenberg, Raithel, and Harbecke 2019).

Finally, Utsumi (2018, 2020) carried out an analysis of the semantic content of non-contextualized word embeddings, which is close in spirit to our correlation analyses in Section 4. However, our study significantly differs from Utsumi's for its goals and scope. Whereas Utsumi (2020) only aims at understanding the semantic knowledge encoded in distributional vectors, we add to this the idea of using the decoded embeddings to explain and to interpret their performance in probing semantic tasks (Section 5). Moreover, our study involves a larger array of DSMs and it is the first one to include state-of-the-art contextualized embeddings.[1]

3. Decoding the Semantic Content of Word Embeddings

We decode the meaning of word embeddings $\mathbf{e_1}, \ldots, \mathbf{e_n}$ by mapping them onto vectors of human-interpretable semantic features $\mathbf{f_1}, \ldots, \mathbf{f_n}$. We henceforth use the term *dimension* to refer to embedding components, and we reserve the term *(semantic) feature* only for interpretable meaning components. First, we present the DSMs we have used in our experiments (Section 3.1), then we introduce the Binder features (Section 3.2), we illustrate the mapping method (Section 3.3), and we evaluate its quality (Section 3.4).

1 After the article submission, we learned about the contribution by Turton, Vinson, and Smith (2020), who also presented a set of regression experiments to map word embeddings onto the Binder feature space. Similarly to Utsumi (2020), however, their analysis is limited to traditional, non-contextualized embedding models.

3.1 Word Embedding Models

Because we aim at providing a systematic comparison of the most common DSMs, we evaluate a large pool of standard, non-contextualized word embedding models. We trained 300-dimensional vectors on a corpus of about 3.9 billion tokens, obtained from the concatenation of ukWaC (Baroni et al. 2009) and a 2018 dump of Wikipedia. All vectors share the same vocabulary of ca. 345K unlemmatized tokens, corresponding to the words with a minimum frequency of 100 in the training corpus.

Our "model zoo" includes both predict models (**SGNS** and **FastText**) and count models (**PPMI** and **GloVe**). SGNS (Mikolov et al. 2013) is the Skip-Gram with Negative Sampling algorithm, which learns word embeddings that predict the context lexemes co-occurring with the targets. FastText (Bojanowski et al. 2017) is a variation of SGNS that uses subword information and represents word types as the sum of their *n*-gram embeddings. GloVe (Pennington, Socher, and Manning 2014) is a matrix model that uses weighted regression to find embeddings that minimize the squared error of the ratios of co-occurrence probabilities. PPMI (Bullinaria and Levy 2012) consists of a co-occurrence matrix weighted with Positive Pointwise Mutual Information and reduced with SVD. Although the latter DSM type could be considered out of date, we decided to include it in our experiments, since Levy, Goldberg, and Dagan (2015) have shown that it can be competitive with predict models, given a proper hyperparameter optimization.

Four DSMs are window-based (the **w2** models select co-occurrences with a context window of 2 words to either side of the target), and four are syntax-based. The **synt** models use contexts typed with syntactic dependencies (e.g., *eat-nobj*), while the **synf** models use syntactically filtered, untyped contexts. Dependencies were extracted from the training corpus parsed with CoreNLP (Manning et al. 2014). As suggested by Levy, Goldberg, and Dagan (2015) for the parameter tuning of count models, we used the context distribution smoothing of 0.75 for PPMI and we dropped the singular value matrix produced by SVD. We also applied to PPMI and GloVe the subsampling method proposed in Mikolov et al. (2013). A summary of all the models with their respective training hyperparameters is provided in Table 1.

The contextualized embedding models are **ELMo**[2] and **BERT** (the BERT-Large uncased version).[3] Because they produce token vectors, following the method proposed by Bommasani, Davis, and Cardie (2020) and Vulić et al. (2020), we created type representations by randomly sampling 1,000 sentences for each target word from the Wikipedia corpus. We generated a contextualized embedding for each word token by feeding the sentence to the publicly available pre-trained models of ELMo and BERT and taking the token vector of the output layer. Finally, an embedding for each word was obtained by averaging its 1,000 contextualized vectors. Averaging contextualized embeddings has been shown to produce vectors that are competitive or even better than those produced by static DSMs. Moreover, this choice is consistent with the hypothesis that context-independent conceptual representations are abstractions from token exemplar concepts (Yee and Thompson-Schill 2016).[4] As a baseline, we also built models based on 300-dimensional randomly generated vectors (**Random**).

2 `https://tfhub.dev/google/elmo/3`.

3 We used the pipelines included in the spacy-transformers package (`https://spacy.io/universe/project/spacy-transformers`).

4 One reviewer correctly points out that we could have queried ELMO and BERT with the very same sentences used by Binder et al. (2016) to clarify the relevant word sense to the workers in the rating tasks (cf. Section 3.2). Unfortunately, these sentences were not released together with the data set.

Table 1
List of the embedding models used for the study, together with their hyperparameter settings.

Model	Hyperparameters
PPMI.w2	345K window-selected context words, window of width 2 weighted with Positive Pointwise Mutual Information (PPMI) reduced with Singular Value Decomposition (SVD) subsampling method from Mikolov et al. (2013).
PPMI.synf	345K syntactically filtered context words weighted with Positive Pointwise Mutual Information (PPMI) reduced with Singular Value Decomposition (SVD) subsampling method from Mikolov et al. (2013).
PPMI.synt	345K syntactically typed context words weighted with Positive Pointwise Mutual Information (PPMI) reduced with Singular Value Decomposition (SVD) subsampling method from Mikolov et al. (2013).
GloVe	Window of width 2 subsampling method from Mikolov et al. (2013).
SGNS.w2	Skip-gram with negative sampling window of width 2, 15 negative examples trained with the `word2vec` library (Mikolov et al. 2013).
SGNS.synf	Skip-gram with negative sampling syntactically-filtered context words, 15 negative examples trained with the `word2vecf` library (Levy and Goldberg 2014).
SGNS.synt	Skip-gram with negative sampling syntactically-typed context words, 15 negative examples trained with the `word2vecf` library (Levy and Goldberg 2014).
FastText	Skip-gram with negative sampling and subword information window of width 2, 15 negative examples trained with the `fasttext` library (Bojanowski et al. 2017).
ELMo	Pretrained ELMo embeddings (Peters et al. 2018), available at `https://allennlp.org/elmo`, original model trained on the 1 Billion Word Benchmark (Chelba et al. 2013).
BERT	Pretrained BERT-Large embeddings (Devlin et al. 2019) available at `https://github.com/google-research/bert` model trained on the concatenation of the Books corpus (Zhu et al. 2015) and the English Wikipedia.

3.2 The Binder Data Set: Features for Brain-Based Semantics

Binder et al. (2016) proposed a **brain-based semantics** consisting of conceptual primitives defined in terms of the *modalities of neural information processing*. This study aimed at developing a representation that captured aspects of experience that are central in the acquisition of concepts. The authors organized human experience in 14 different domains (see Table 2), each one corresponding to a variable number of features for which some specialized neural processor has been identified and described in the

Table 2
List of the domains and meaning components (features) in Binder et al. (2016).

Domain	Meaning components (features)
Vision	VISION, BRIGHT, DARK, COLOUR, PATTERN, LARGE, SMALL, MOTION, BIOMOTION, FAST, SLOW, SHAPE, COMPLEXITY, FACE, BODY
Somatic	TOUCH, TEMPERATURE, TEXTURE, WEIGHT, PAIN
Audition	AUDITION, LOUD, LOW, HIGH, SOUND, MUSIC, SPEECH
Gustation	TASTE
Olfaction	SMELL
Motor	HEAD,UPPER LIMB, LOWER LIMB, PRACTICE
Spatial	LANDMARK, PATH, SCENE, NEAR, TOWARD, AWAY, NUMBER
Temporal	TIME, DURATION, LONG, SHORT
Causal	CAUSED, CONSEQUENTIAL
Social	SOCIAL, HUMAN, COMMUNICATION, SELF
Cognition	COGNITION
Emotion	BENEFIT, HARM, PLEASANT, UNPLEASANT, HAPPY, SAD, ANGRY, DISGUSTED, FEARFUL, SURPRISED
Drive	DRIVE, NEEDS
Attention	ATTENTION, AROUSAL

neuroscientific literature (Binder et al. 2016). In total, the brain-based semantics consists of 65 cognitively motivated features, which we henceforth refer to as the **Binder features**.

For their collection of ratings, Binder et al. (2016) selected 242 words of the *Knowledge Representation in Neural Systems project* (Glasgow et al. 2016), including 141 nouns, 62 verbs, and 39 adjectives, plus 293 additional nouns in order to include more abstract nouns, for a total of 535 words. Rated words belong to various concept types. A summary of the concept types, parts-of-speech, and the number of words per type is provided in Table 4. For each of these words, ratings on a 0–6 scale were collected with Amazon Mechanical Turk, in order to assess the degree to which humans associate their meaning with particular kinds of experience. Words were rated across multiple sessions: Each participant was assigned one word per session and provided ratings for all the semantic features (cf. Table 3 for an example). Because there are several ambiguous words in the data, participants were presented with an example sentence that allowed the correct identification of the target word sense. The reported mean intraword individual-to-group correlation of the collected ratings is 0.78 (median 0.80). Interestingly, the concept representations based on the elicited features were compared with their distributional representations, obtained via Latent Semantic Analysis (Landauer and Dumais 1997), showing that brain-based features are more efficient in separating conceptual categories.

Table 3
A sample of the rated Binder features for *dog* and *love*.

Word	VISION	BRIGHT	...	COGNITION	BENEFIT	HARM	PLEASANT	UNPLEASANT
dog	5.3548	1.0968	...	0.3548	3.5806	2.8065	3.9355	0.7097
love	0.7931	0.4828	...	4.5172	4.9310	1.7586	5.4828	0.5172

Table 4
Concept types, parts-of-speech (POS), and number of items in the data set by Binder et al. (2016).

Type-POS	No. of items
Concrete Objects - Nouns	275
Living Things - Nouns	126
Other Natural Objects - Nouns	19
Artifacts - Nouns	130
Concrete Events - Nouns	60
Abstract Entities - Nouns	99
Concrete Actions - Verbs	52
Abstract Actions - Verbs	5
States - Verbs	5
Abstract Properties - Adjectives	13
Physical Properties - Adjectives	26

We have chosen the Binder features for our decoding experiments for three main reasons. First of all, they are empirically motivated on the grounds of neurocognitive evidence supporting their key role for conceptual organization. This allows us to test the extent to which these central components of meaning are actually captured by word embeddings. Second, despite being quite coarse-grained, Binder features differ from human generated properties because the latter are actually linguistic structures that often express complex concepts (e.g., *used_for_transportation* as a property for *airplane*), rather than core meaning components. Third, the Binder data set covers nouns, verbs, and adjectives, and encompasses both concrete and abstract words, while no existing feature norms have a comparable morphosyntactic or semantic variety. Of course, we do not claim this to be the "definitive" semantic feature lists, but in our view it represents the most complete repository of continuous featural representations available to date. However, the analysis methodology we present in the next section is totally general, and can be applied to any type of semantic feature vector.

3.3 Mapping Word Embeddings onto Semantic Features

For this study, we learn a mapping from an n-dimensional word embedding $\mathbf{e}$ (with n equal to 300 for non-contextualized DSMs, 768 for BERT, and 1,024 for ELMo) onto a 65-dimensional feature vector $\mathbf{f}$ whose components correspond to the ratings for the Binder features. We henceforth refer to the mapped feature vectors as **Binder vectors**. Our data set consists of 534 Binder words.[5]

In the previous literature, mainly two methods have been used to learn a mapping between embeddings and discrete feature spaces: regression models (Făgărășan, Vecchi, and Clark 2015; Pereira et al. 2018; Utsumi 2018, 2020) and feedforward neural networks (Abnar et al. 2018; Utsumi 2018, 2020). We performed our experiments with a partial least square regression model, with an appropriately chosen number k of regression components. We tested with $k = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$. The regression models have been implemented with the Python Scikit-learn package (Pedregosa et al.

5 One less than the original collection, because *used* appears twice, as verb and adjective.

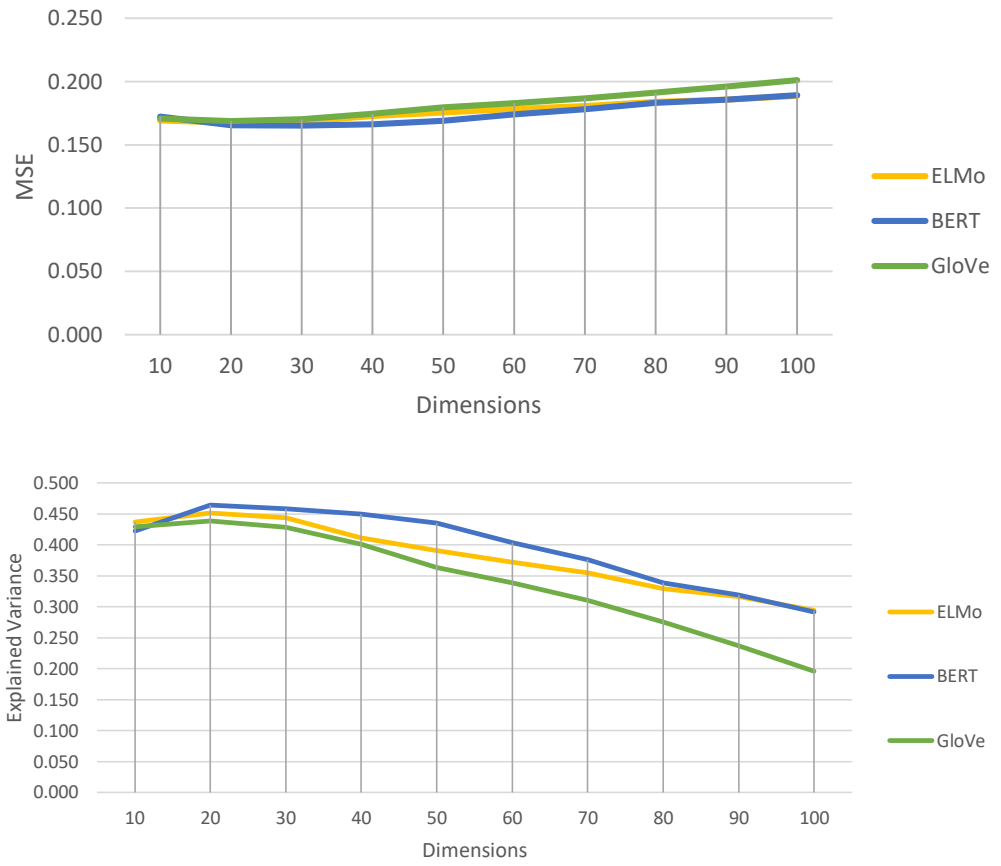


Figure 1
(top) Mean Squared Error (values have been summed across the Binder features) and (bottom) explained variance for ELMo, BERT, and GloVe vectors per number of regression components.

2011).[6] Identifying the best mapping model is not the main goal of the present work, and we leave for further research the comparison with other methods.

3.4 Mapping Evaluation

For a preliminary evaluation of the mapping quality, we analyze the traditional metrics of Mean Squared Error (*MSE*) and variance. First, we randomly split the data into training and test sets, using an 80 : 20 ratio, and we measure the sum of *MSE* and the variance in order to determine the optimal value for the parameter k (the number of regression components). After choosing the optimal k, vectors of Binder features are predicted with the leave-one-out training paradigm, as in Utsumi (2018, 2020): For each word in the data set, we train a mapping between the embeddings and the gold

6 https://scikit-learn.org/stable/.

Table 5
Mean Squared Error (summed across features) and explained Variance per model, mapping with Partial Least Squares Regression and $k = 30$.

Model	MSE	Variance
PPMI.w2	0.16	0.50
PPMI.synf	0.15	0.54
PPMI.synt	0.16	0.48
GloVe	0.16	0.46
SGNS.w2	0.15	0.54
SGNS.synf	**0.14**	0.58
SGNS.synt	**0.14**	0.59
FastText	**0.14**	0.55
ELMo	0.16	0.48
BERT	0.15	0.53
Random	0.30	−0.73

standard Binder vectors of the remaining words, and we predict the held-out word with the resulting mapping.

Moreover, following the literature on neural decoding, the predicted vectors are tested on a task of *retrieval in the top-N neighbors*. Given a predicted vector, we rank all the 534 vectors in the gold standard data set by decreasing cosine similarity values. Then we measure the Top-N accuracy (*Top-N Acc*), as the percentage of the items of the data set whose gold standard vector is in the top-N of the neighbors list of the predicted vector (Făgărășan, Vecchi, and Clark 2015). We assess this value for $N = 1, 5, 10$.

We measured the *MSE* and the explained variance for each model, finding that k between 30 and 50 produces the optimal results for all models. Figure 1 shows *MSE* and variance as a function of k for GloVe, ELMo, and BERT embeddings. Most models achieve the best fit with $k = 30$ or $k = 40$. Since the average explained variance is slightly higher for $k = 30$, we keep this parameter as the optimal value for the mapping. Table 5 reports the *MSE* and the explained variance for the $k = 30$ mapping. The best scores are obtained with the syntax-based versions of the SGNS model, together with FastText and BERT. All mappings perform largely better than the random baseline, for which the explained variance is negative.

Using the Partial Least Squares Regression model with $k = 30$ for the mapping and leave-one-out training, we predict the vectors of all the Binder words and we evaluate them with the Retrieval task. The results are shown in Table 6. At a glance, we can notice that all DSMs vastly outperform the random vectors and are able to retrieve in the top 10 ranks at least half of the target concepts. For this auxiliary task, the best performing model is BERT, which retrieves the 30% of the target concepts at the top of the ranking and more than three quarters of them in the top 10. This confirms that type vectors derived from contextualized embeddings can indeed be competitive with those generated by static DSMs (Chronis and Erk 2020; Vulić et al. 2020). The next best models are the syntactically enriched versions of the Skip-Gram vector, with the one using typed dependencies coming close to BERT performance.

These results show overall good-quality representation for all the embedding types, and a comparison with the scores by Utsumi (2018) confirms the superiority of SGNS model over GloVe and PPMI for this kind of mapping. Differently from the previous study, we also consider embeddings that are trained with syntactic dependencies,

Table 6
Top-N Acc for each word embedding model.

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

showing that for SGNS syntactic contexts determine a general improvement of the performance (while typed dependencies are suboptimal for the PPMI model).

The next tests will aim at revealing how well the different features in the Binder data set are encoded by our vectors.

4. How Do Word Embeddings Encode Semantic Features?

In the literature on neurosemantic decoding, it has been shown that models can be compared for their ability to explain the activity in a particular brain region (Wehbe et al. 2014; Gauthier and Ivanova 2018; Anderson et al. 2018). Analogously, we want to inspect which features are better predicted by a given embedding model. We compute the average of the Spearman correlation between human ratings and model predictions across words and features. Results are reported in Table 7.

Table 7
Average word and feature Spearman correlation between the human ratings in Binder et al. (2016) and the estimated values for each embedding type.

Model	Word correlation	Feature correlation
PPMI.w2	0.77	0.70
PPMI.synf	0.79	0.72
PPMI.synt	0.80	0.71
GloVe	0.76	0.69
SGNS.w2	0.80	0.74
SGNS.synf	**0.82**	0.76
SGNS.synt	**0.82**	**0.77**
FastText	0.81	0.75
ELMo	0.81	0.76
BERT	**0.82**	**0.77**
Random	0.20	−0.01

All DSMs achieve high correlation values, higher than 0.7 per word, vastly outperforming the random baseline. The results across the models are similar, again with BERT and the syntax-based SGNS models taking the top spots. Consistently with the previous tests, syntactic information seems to be useful for predicting the property values, as syntax-based models almost always perform slightly better than their window-based counterparts. A similar finding for the prediction of brain activation patterns has already been descibed by Abnar et al. (2018), who also reported a strong performance by dependency-based embeddings. It is also interesting to notice that all our models have much higher correlation values than the best results reported by Utsumi (2018), a difference that might be due to the choice of the training corpora (we used a concatenation of Ukwac and Wikipedia, whereas Utsumi trained his models on the COCA and, separately, on the Wikipedia corpus alone). Finally, while the PPMI embeddings used by Utsumi drastically underperform, our PPMI vectors come much closer to the predict ones, although the latter still retain an edge.

In the heatmap in Figure 2, it is possible to observe the average correlations per Binder domain. It is striking that the features belonging to the Cognition, Causal, and Social domains are the best predicted ones, together with the Gustation domain, which however includes just one feature. On the other hand, other somatosensorial domains are predicted with lower accuracies. As suggested by Utsumi (2018, 2020), who reported consistent findings, this can be explained by the fact that embeddings learn word meaning only from textual data: Psycholinguistic studies on mental lexicon theorize that humans combine both linguistic information and first-hand experience of the world (Vigliocco and Vinson 2007; McRae and Matsuki 2009), and domains such as Cognition and Social are especially important in the characterization of abstract concepts, for which textual information has been suggested to be the prevailing source (Vigliocco et al. 2009). When it comes to the somatosensorial features of concrete concepts, instead,

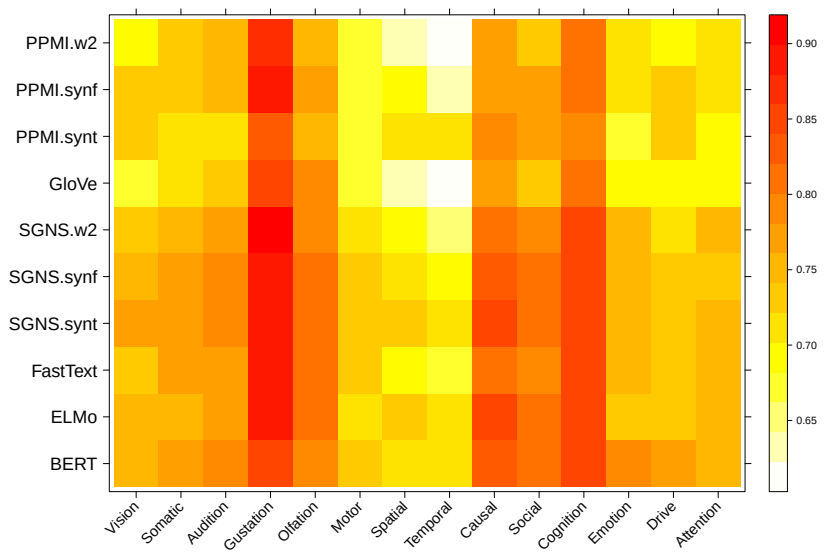


Figure 2
Average Spearman correlations per domain between the estimated and the original Binder features for each embedding type.

text-based models are clearly missing that kind of information on the referents, although various aspects of experiential information are "redundantly" encoded in linguistic expressions (Riordan and Jones 2011), as proposed by the so-called *Symbol Interdependency Hypothesis* (Louwerse 2008).

Finally, spatial and temporal features are particularly challenging for distributional representations. This is compatible with the hypothesis that temporal concepts are mainly represented in spatial terms and the acquisition of spatial attributes requires multimodal evidence (Binder et al. 2016), which is instead lacking in our distributional embeddings. The Emotion domain also shows good correlation values, confirming the role of distributional information in shaping the affective content of lexical items (Recchia and Louwerse 2015; Lenci, Lebani, and Passaro 2018).

Figure 3 provides a more analytical and variegated view of the way embeddings predict each Binder feature, revealing interesting differences within the various domains. First of all, we can observe that some somatosensorial semantic dimensions are indeed strongly captured by embeddings, consistent with the hypothesis that several embodied features are encoded in language (Louwerse 2008). For instance, COLOR (i.e., "having a characteristic or defining color"), MOTION (i.e., "showing a lot of visually observable movement"), BIOMOTION (i.e., "showing movement like that of a living

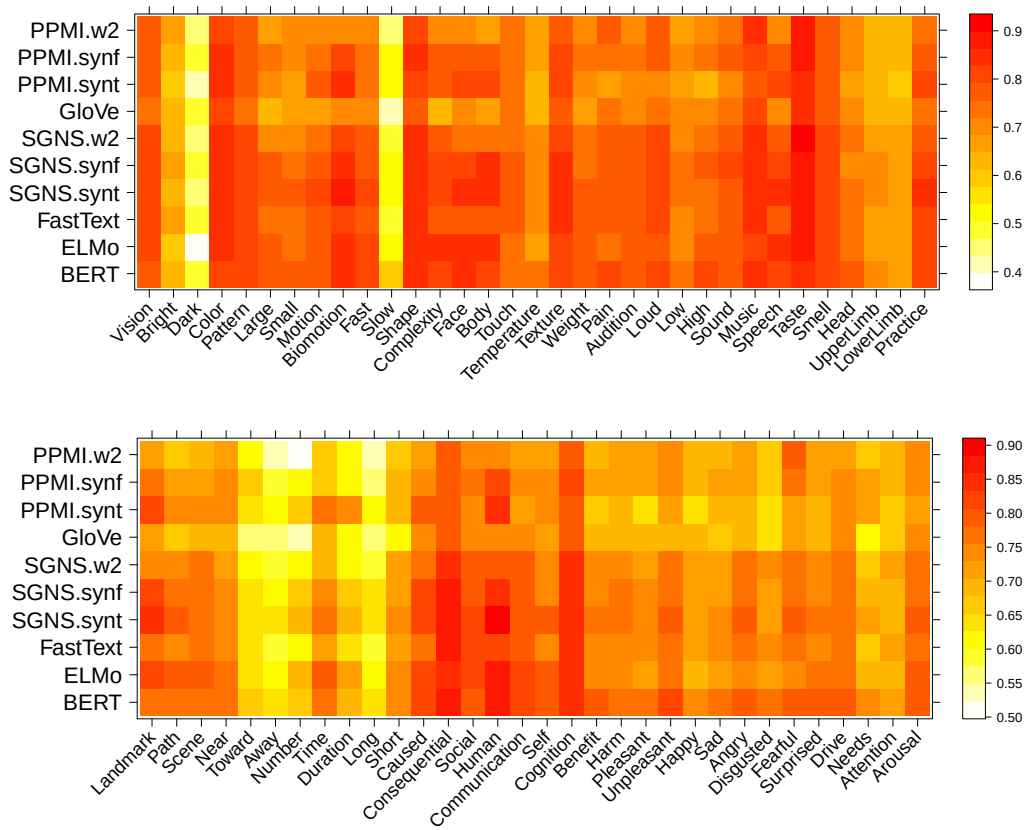
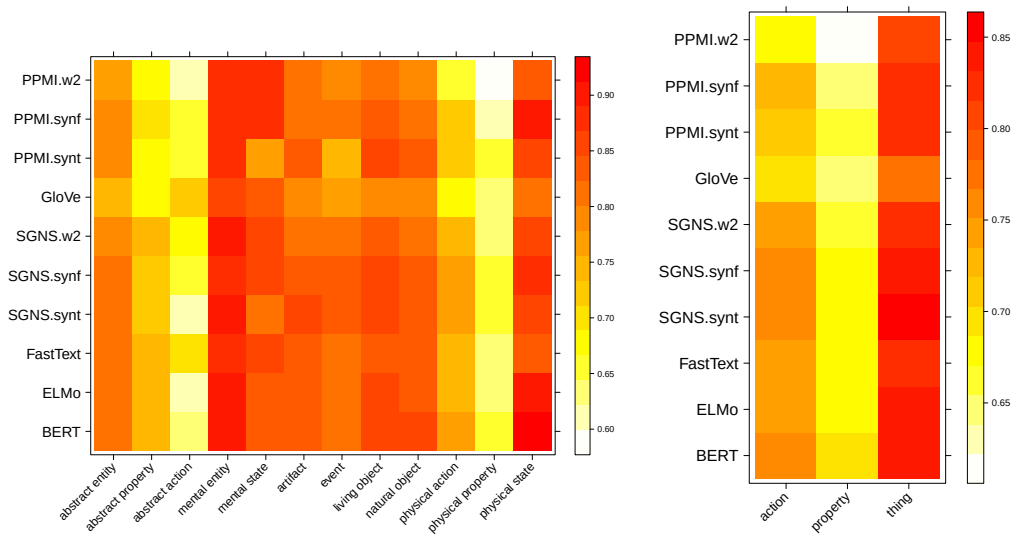


Figure 3
Spearman correlations between the estimated and original Binder features for each embedding type.

thing"), and SHAPE (i.e., "having a characteristic or defining visual shape or form") are among the best predicted visual features. FAST (i.e., "showing visible movement that is fast") is predicted much better than SLOW (i.e., "showing visible movement that is slow"), while embeddings do not seem to discriminate between the BRIGHT (i.e.,"visually light or bright") and DARK (i.e., "visually dark") components.

In the Audition domain, LOUD (i.e., "making a loud sound"), MUSIC (i.e., "making a musical sound"), and SPEECH (i.e., "someone or something that talks") are generally very well predicted. The Spatial domain instead shows an uneven behavior, with LANDMARK (i.e., "having a fixed location, as on a map"), SCENE (i.e., "bringing to mind a particular setting or physical location"), and PATH (i.e., "showing changes in location along a particular direction or path") presenting much higher correlation values than the other features. The best predicted social features are HUMAN (i.e., "having human or human-like intentions, plans, or goals") and COMMUNICATION (i.e., "a thing or action that people use to communicate"). In relation to spatial features, TIME (i.e., "an event or occurrence that occurs at a typical or predictable time") and HUMAN, it is interesting to point out that the DSMs with syntactic information generally produce better predictions than their window-based equivalents (cf. in Figure 3 the values for the synf/synt versions of PPMI/SGNS models with their w2 equivalents and with FastText). Finally, negative sentiments and emotions are better predicted than positive ones. This is consistent with previous reports of negative moods being more frequently expressed in online texts (De Choudhury, Counts, and Gamon 2012).

Using the metadata in the Binder data set, we group the words per super category and type, and compute the average correlations. A quick look at Figure 4a reveals that mental entities are the best represented ones, while embeddings struggle the most with physical and abstract properties. Also note that living objects and events tend to be well represented by most embedding models. Figure 4b provides a summary of the average correlations per word type, confirming that things are the most correlated, whereas



(a) (b)

Figure 4
Average Spearman correlations per word super category (a) and per word super type (b).

weaker correlations are observed for actions. Only moderate-to-low correlations are achieved for properties.

Finally, it is worth focusing on the behavior of contextualized embeddings. Though BERT has a slightly higher Top-N accuracy (cf. Section 3.4), its overall word and feature correlation is equivalent to the one by SGNS.synt (cf. Table 7). Moreover, figures 2–4 do not show any significant difference in the kinds of semantic dimensions encoded by traditional DSMs with respect to BERT and ELMo vectors. This leads us to conjecture that the true added value of the latter models lies in their ability to capture the meaning variations of word tokens in context, rather than in the type of semantic information they can distill from distributional data.

5. Using Semantic Features to Analyze Probing Tasks

Probing tasks (Ettinger, Elgohary, and Resnik 2016; Adi et al. 2017) have become one of the most common tools to investigate the content of embeddings. A probing task consists of a binary classifier that is fed with embeddings and is trained to classify them with respect to a certain linguistic dimension (e.g., animacy). The classification accuracy is taken as proof that the embeddings encode that piece of linguistic information. As we have said in Section 1, the limit of the probing task methodology is that it only provides *indirect* evidence about the way linguistic categories are represented by embeddings. In this section, we show how the decoded Binder vectors can be used to "open the box" of semantic probing tasks, to *inspect* the features that are relevant for a certain task, and to *analyze* the performance of distributional embeddings.

5.1 Probing Tasks for Human-Interpretable Embeddings

We use the original word embeddings and their corresponding mapped Binder vectors as input features to a logistic regression classifier, which has to determine if they belong to a given semantic class. The human-interpretable nature of Binder vectors allows us to decode and explain the performance of the original embeddings in the probing tasks.

Being able to determine the semantic class of a word is an important cognitive and linguistic task (Murphy 2002). Research on the automatic identification of semantic classes is central in computational linguistics (Vulić et al. 2017; Yaghoobzadeh et al. 2019). The detection of a given semantic feature of a word is potentially useful for the automatic creation of lexicon and dictionaries, such as sensory lexicons (Tekiroglu, Özbal, and Strapparava 2014), emotion lexicons (Buechel and Hahn 2018), and sentiment dictionaries (Turney and Littman 2003; Esuli and Sebastiani 2006; Baccianella, Esuli, and Sebastiani 2010; Cardoso and Roy 2016; Sedinkina, Breitkopf, and Schütze 2019). Linguistic research also benefits from automatic methods to classify linguistic expressions according to various semantic dimensions (Boleda 2020).

Non-contextualized embeddings were tested on the following probing tasks that target different semantic classes and features:

> **Positive/Negative** – Given the embedding of a word, the classifier has to decide whether the word has a positive or a negative polarity. The data set consists of 250 positive words and 250 negative words from the ANEW sentiment lexicon (Nielsen 2011; Bradley and Lang 2017), which is composed of a total of 3,188 words with human valence ratings on a scale between 1 (very unpleasant) and 8 (very pleasant). The selected positive items have valency ratings higher than 7, and the negative items have

valency ratings lower than 3. The data set was randomly split in 400 items for training and 100 words for test.

Concrete/Abstract – The task is to decide whether a noun is concrete or abstract. The data set consists of 254 nouns (91 abstract, 163 concrete) selected from SimLex-999 (Hill, Reichart, and Korhonen 2015). The data set was randomly split in 203 items for training and 51 words for test. For this task, concrete nouns are assumed as the positive class.

Animate/Inanimate – The task is to decide whether a noun is animate or inanimate. The data set includes 810 nouns (672 animate, 138 inanimate) corresponding to the targets in the Direct object animacy task described below, randomly split in 658 for training and the remaining 152 for test. For this task, animate nouns are assumed as the positive class.

VerbNet – The task is based on verb semantic classes included in VerbNet (Kipper et al. 2008; Palmer, Bonial, and Hwang 2017). For each VerbNet class, we generated a set of negative examples by randomly extracting an equal number of verbs that do not belong to the semantic class (i.e., for a semantic class with n verbs, we extract n verbs from the other classes to build the negative examples). Each class was then randomly split in a training and in a test set, using an 80:20 ratio, and we selected the 23 classes that contained at least 20 test verbs.[7] The task consists of predicting whether a target verb is a class instance or not.

As the key feature of contextualized DSMs is to generate embeddings of words in context, BERT and ELMo were tested on two semantic tasks probing a target word token in an input sentence:

Direct object animacy – The task is to decide whether the direct object noun of a sentence is animate or inanimate, and is the contextualized equivalent of the Animate/Inanimate task above. The data set includes 647 training subject - verb - object sentences with animate and inanimate direct objects, and 163 test sentences.[8]

Causative/Inchoative alternation – The task is to decide whether the verb in a sentence undergoes the causative/inchoative alternation or not (Levin 1993). Alternating verbs like *break* can occur both in agent-patient transitive sentences (*The man broke the vase*) and in intransitive sentences in which the patient noun occurs as subject (*The vase broke*). Non-alternating verbs like *buy* can instead only occur in transitive sentences (*The man bought the book* vs. **The book bought*). This task has already been used to probe vectors by Warstadt, Singh, and Bowman (2019) and Klafka and Ettinger (2020). We used the data set of the latter work, consisting of 4,000 training sentences and 1,000 test sentences, equally split between alternating and non-alternating target verbs. For this task, alternating verbs are assumed as the positive class.

7 The classes pour-9.5+spray-9.7, remove-10.1+clear-10.3+mine-10.9, and cut-21.1+carve-21.2 were obtained by merging some VerbNet subclasses.

8 The data set was developed and kindly provided by Evelina Fedorenko, Anna Ivanova, and Carina Kauf.

BERT and ELMo were queried with the sentences in the data set to obtain the contextualized embedding of the target word (the direct object noun for the animacy task, the verb for the causative/inchoative one), which was then fed into the classifiers.

The embeddings were not fine-tuned in the probing tasks. In fact, the overall purpose of the analysis is not to optimize the performance of the classifiers, but to use them to investigate the information that the original embeddings encode.

5.2 Interpreting Probing Tasks with Binder Features

Our analysis consists of three main steps: (i) for each semantic task, we first train a classifier using the original word embeddings and we measure their accuracy, as customary in the probing task literature; (ii) then, we train the same classifiers using the corresponding mapped Binder vectors in the training sets and we inspect the most important semantic features of each probed class; finally, (iii) we measure the overlap between the classifier top features and the top features of the words in the test sets, and we use this information to interpret the performance of the models in the various tasks.

5.2.1 Measuring Embedding Accuracy in Probing Tasks. First of all, we evaluate the performance of the embeddings in each task via the traditional accuracy metric, in order to check their ability to predict the semantic class of the word. A summary of the performance of the traditional DSMs can be seen in Table 8, while the scores for the contextualized models are shown in Table 9. Because the classes are unbalanced in most tasks, the tables also report the results for a majority baseline. At a first glance, in Table 8 we can notice a performance gap between count models based on PPMI and SVD and the other word embedding models, with the former being largely outperformed by the latter on all probing data sets (the largest observed gap being around 40%) and struggling even to beat the majority baseline in many of the VerbNet-derived test sets. All neural embeddings achieve a 100% accuracy in the classification for the Concrete/Abstract task and one of the models, FastText, achieves the same score also on the Positive/Negative task. The VerbNet tasks, possibly because of the fuzzy boundaries of the verb semantic classes, proved to be the most challenging ones and in some cases the models struggle to beat a chance classifier. The best performing embeddings are, in general, the FastText ones and the vectors of the SGNS family.

As for the contextualized probing tasks, BERT outperforms ELMo, and the Causative/Inchoative alternation task is more difficult, probably because alternating verbs are semantically more heterogeneous. However, even in this case the classification accuracy is very high, when compared to the majority baseline.

5.2.2 Examining the Semantic Features of the Probed Classes. Because probing tasks are typically used as "black box" tools, the performance obtained by a certain DSM is usually regarded to be enough to draw conclusions about the information encoded by its vectors. Here, the mere embedding accuracy we have reported in Tables 8 and 9 is not the primary aim of our analyses. In fact, we want to make the semantic information learned by the classifiers explicit and human-interpretable, in order to characterize the content of the probed semantic dimensions. To this purpose:

- for each DSM, we learn a mapping between its embeddings and the 65-dimensional Binder vectors, using the whole set of 534 Binder words as training data;

Table 8
Classification accuracy on the probing tasks for the 8 non-contextualized DSMs.

Task	PPMI			SGNS			GloVe	FastText	Majority
	w2	synf	synt	w2	synf	synt			
Positive/Negative	0.52	0.65	0.65	0.83	0.85	0.79	0.79	**1.00**	0.52
Concrete/Abstract	0.69	0.70	0.73	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.59
Animate/Inanimate	0.74	0.78	0.86	0.94	0.96	0.97	0.96	**0.98**	0.83
VerbNet									
pour-9.5+spray-9.7	0.61	0.52	0.48	0.70	0.78	0.78	**0.82**	0.74	0.52
fill-9.8	0.50	0.50	0.52	**0.74**	0.72	**0.74**	0.70	**0.74**	0.50
butter-9.9	0.47	0.55	0.50	0.87	0.84	0.86	**0.89**	**0.89**	0.63
pocket-9.10	0.58	0.50	0.50	0.88	0.88	**0.92**	0.75	0.83	0.50
remove-10.1+clear-10.3 +mine-10.9	0.44	0.56	0.44	0.56	0.68	**0.76**	0.68	0.64	0.52
steal-10.5	0.48	0.48	0.41	0.83	0.83	0.79	0.86	**0.90**	0.52
debone-10.8	0.59	0.63	0.54	0.86	0.82	0.82	0.68	**0.90**	0.50
cut-21.1+carve-21.2	0.65	0.65	0.57	0.80	0.80	**0.96**	0.57	0.73	0.50
amalgamate-22.2	0.56	0.52	0.52	0.74	0.70	0.78	0.74	**0.83**	0.52
tape-22.4	0.65	0.62	0.73	0.98	0.97	0.94	0.98	**1.00**	0.59
characterize-29.2	0.57	0.57	0.57	0.81	0.81	**0.86**	0.76	0.81	0.52
amuse-31.1	0.55	0.56	0.51	0.69	0.80	**0.75**	0.67	0.72	0.51
admire-31.2	0.44	0.52	0.56	0.87	**0.91**	0.87	0.78	0.96	0.52
marvel-31.3	0.59	0.62	0.65	0.72	0.69	**0.83**	0.69	0.76	0.58
judgement-33.1	0.71	0.66	0.66	0.77	**0.80**	**0.80**	0.77	**0.80**	0.54
manner_of_speaking-37.3	0.71	0.79	0.48	0.79	0.86	**0.90**	**0.90**	0.86	0.50
say-37.7	0.62	0.60	0.50	0.55	0.50	**0.64**	0.50	0.55	0.55
animal_sounds-38	0.67	0.80	0.67	0.87	0.90	0.83	0.83	**0.93**	0.56
sound_emission-43.2	0.52	0.48	0.61	0.70	**0.78**	0.74	0.65	0.70	0.52
cooking-45.3	0.55	0.60	0.45	0.77	0.86	**0.90**	0.73	0.86	0.52
other_cos-45.4	0.54	0.50	0.57	0.70	0.73	0.74	**0.76**	0.68	0.50
contiguous_location-47.8	0.57	0.57	0.52	**0.86**	0.76	0.81	0.61	0.81	0.52
run-51.3.2	0.52	0.48	0.56	0.80	**0.84**	0.80	0.75	0.80	0.56

Table 9
Classification accuracy on the contextualized probing tasks.

Task	BERT	ELMo	Majority
Direct object animacy	**0.99**	0.96	0.83
Causative/Inchoative alternation	**0.91**	0.86	0.51

- we use the decoding mapping to generate the Binder vectors of all the words contained in the probing data sets;
- for each probing task, we train a classifier t with the decoded Binder vectors;
- we extract the weights assigned by the classifier to the Binder features and sort them in descending order. Given the task t, $TopTaskFeats(t, n)$ is the set of the top n features learned by the classifier for t using the Binder vectors.

Table 10
Top 5 features ordered from left to right for a selection of the non-contextualized probing tasks with the Binder vectors mapped from the FastText embeddings, and for the contextualized probing tasks with the Binder vectors mapped from the BERT token embeddings.

Task	Top Features
Positive/Negative	PLEASANT, HAPPY, BENEFIT, NEEDS, SELF
Concrete/Abstract	SHAPE, VISION, WEIGHT, TEXTURE, TOUCH
Animate/Inanimate	FACE, BODY, HUMAN, SPEECH, BIOMOTION
fill-9.8	COLOR, VISION, BRIGHT, WEIGHT, PATTERN
cut-21.1+carve-21.2	PRACTICE, TOUCH, UPPER LIMB, VISION, SHAPE
admire-31.2	COGNITION, SOCIAL, AROUSAL, HAPPY, PLEASANT
judgement-33.1	COMMUNICATION, SOCIAL, HEAD, COGNITION, AROUSAL
say-37.7	COMMUNICATION, COGNITION, BENEFIT, SOCIAL, SELF
sound_emission-43.2	AUDITION, LOUD, SOUND, HIGH, MUSIC
cooking-45.3	TASTE, TEMPERATURE, SMELL, HEAD, PRACTICE
contiguous_location-47.8	LANDMARK, VISION, COLOR, LARGE, SCENE
run-51.3.2	LOWER LIMB, MOTION, PATH, FAST, BIOMOTION
Direct object animacy	FACE, UPPERLIMB, SCENE, COMPLEXITY, BIOMOTION
Causative/Inchoative alternation	SLOW, COMPLEXITY, TEMPERATURE, UPPERLIMB, SHORT

The set $TopTaskFeats(t, n)$ includes the most important semantic features for the classification task t. Table 10 reports the top 5 features for some of our probing tasks using the Binder vectors decoded from FastText, which is one of the best performing non-contextualized models on average, and from BERT. Notice that the top features provide a nice characterization of the features of the semantic classes targeted across tasks. FACE, HUMAN, and SPEECH appear among the top features of animate nouns. For sentiment classification, the most relevant features are positive emotions (PLEASANT, HAPPY, BENEFIT) or belong to the Social domain (SELF). On the other hand, physical properties (SHAPE, VISION, WEIGHT) are the most important ones for the Concrete/Abstract distinction, in which concrete nouns represent the positive class. Similar considerations apply for the VerbNet tasks. The class run-51.3.2 contains motion verbs and its most relevant features refer to movement (MOTION, LOWER LIMB, FAST) and direction (PATH). The classes judgment-33.1 and say-37.7 are characterized by features related to communication and cognition. The class sound_emission-43.2 is instead associated with features belonging to the Audition domain. Perhaps the less perspicuous case is represented by the features associated with the alternating class in the Causative/Inchoative task. However, it is worth noticing the salience of the TEMPERATURE feature, as various alternating verbs express this dimension (*warm*, *heat*, *cool*, *burn*, etc.). This shows how a simple featural decoding of the embeddings can be used to investigate the internal structure of the semantic classes that are targeted by probing tasks.

5.2.3 Explaining the Performance of Embeddings in Probing Tasks. The third phase of our analysis combines the results of the previous two steps: The Binder feature vectors learned in Section 5.2.2 are used to explain the accuracy of the embeddings in the probing tasks in Section 5.2.1.

For each task t and word w in the test set of t, we rank the features of the decoded Binder vector $\mathbf{f_w}$ in descending order according to their values. We indicate with $TopWordFeats(\mathbf{f_w}, n)$ the set of the top n features in the Binder vector $\mathbf{f_w}$. Then we

measure with **Average Precision** (*AP*) the extent to which the top Binder features of t appear among the top decoded features of the test word w. Given the ranked feature sets $TopTaskFeats(t, n)$ and $TopWordFeats(\mathbf{f_w}, n)$, we compute $AP(t, w)$ as follows:

$$AP(t, w) = \frac{\sum_1^n P_w^t(r)}{n} \tag{2}$$

$$P_w^t(r) = \frac{|TopTaskFeats(t, n) \cap TopWordFeats(\mathbf{f_w}, n)|_1^r}{r} \tag{3}$$

where the numerator of Equation (3) is the number of task features that are also in the word feature vector from rank 1 to rank r. *AP* is a measure derived from information retrieval combining precision, relevance ranking, and overall recall (Manning, Raghavan, and Schütze 2008; Kotlerman et al. 2010). In our case, the ranked task features are like documents to be retrieved and the word features are like documents returned by a query. *AP* takes into account two main factors: (i) the extent of the intersection among the n most important semantic features for a word and a task, and (ii) their mutual ranking. The higher the $AP(t, w)$ score, the more the top features of w that are also included in the top features for the task t. For example, suppose that $TopTaskFeats(Positive/Negative, 3) = \{\text{PLEASANT}, \text{HAPPY}, \text{BENEFIT}\}$, $AP(Positive/Negative, w) = 1$ if and only if $TopWordFeats(\mathbf{f_w}, 3)$ contains the same semantic features at the top of the rank.

For each model and each task, we analyze the *AP* of the output of the classifiers trained on the original word embeddings, whose accuracy is reported in Tables 8 and 9. We compute the *AP* of the words correctly classified in the positive class (*true positive*, TP) and in the negative class (*true negative*, TN). Moreover, we compute the *AP* of the words wrongly classified in the positive class (*false positive*, FP) and in the negative class (*false negative*, FN). The *AP* distribution of each word group across the probing tasks is reported in Figure 5 for the non-contextualized DSMs and in Figure 6 for BERT and ELMo. The Kruskal-Wallis rank sum non-parametric test shows that in all models the four word groups significantly differ for their *AP* values (df = 3, p-value < 0.001).

Post-hoc pairwise Mann–Whitney U-tests (with Bonferroni correction for multiple comparisons) confirm that across tasks TPs have a significantly higher *AP* than FPs ($p < 0.001$). Therefore, the words correctly classified in the positive class share a large number of the top ranked features for that class (e.g., the words whose embeddings are correctly classified as animate have a large number of the top semantic features that characterize animacy). Conversely, the words correctly classified in the negative class have very few, if any, of the top task features. It is interesting to observe that the DSMs for which the difference between the median *AP* (represented by the thick line in each boxplot) of TPs and the median *AP* of TNs is higher (i.e., the neural embeddings for the non-contextualized models and BERT) are the models that in general show a higher classification accuracy in the probing tasks and better encode the Binder features (cf. Section 4). This suggests that a model accuracy in probing tasks is strongly related to the way its embeddings encode the most important semantic features for a certain classification task (cf. below).

In Figures 5 and 6, the *AP* of the wrongly classified words (i.e., FPs and FNs) tend to occupy an intermediate position between the *AP* of TPs and TNs. In fact, we can conjecture that a word in the positive class (e.g., an animate noun) is wrongly classified (e.g., labeled as inanimate), because it lacks many of the top features characterizing the target class (e.g., animacy). Post-hoc pairwise Mann–Whitney U-tests support this hypothesis, because the *AP* of the FNs is significantly different from the one of TPs

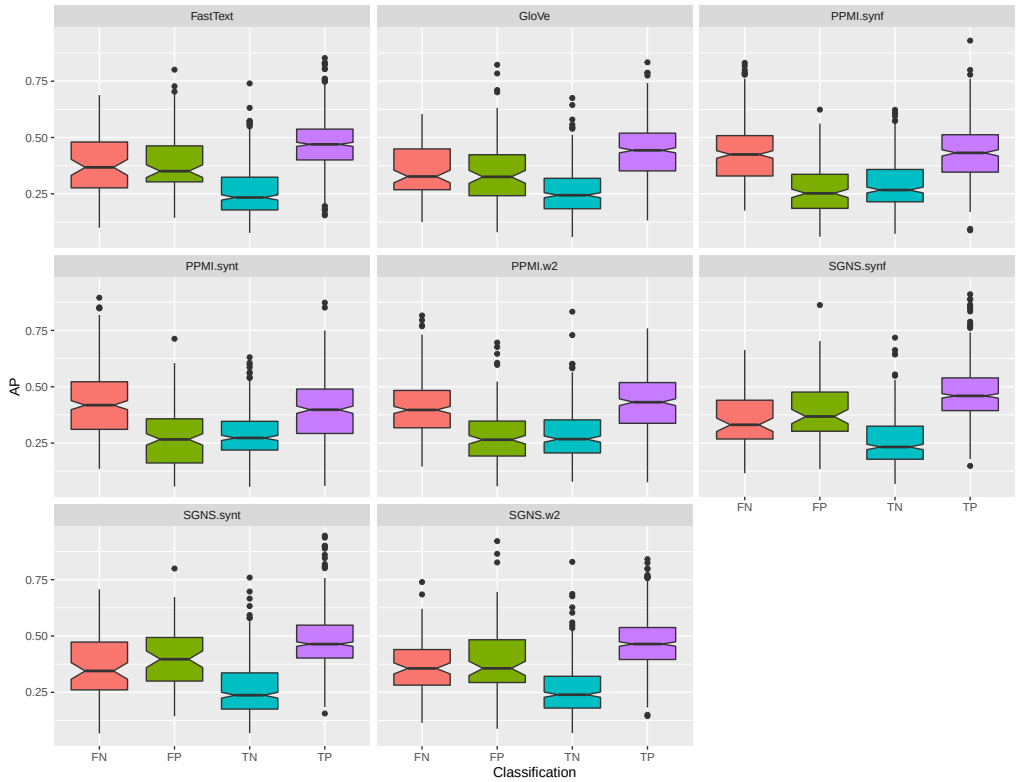


Figure 5
Average Precision (*AP*) boxplots of the Binder vectors of the test words with respect to the top-20 Binder features of each probing task. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) refer to the output of the classifiers trained on the original embeddings of the non-contextualized DSMs.

(PPMI.synt: $p < 0.05$; GloVe: $p < 0.05$; SGNS.w2: $p < 0.001$; SGNS.synf: $p < 0.001$; SGNS.synt: $p < 0.001$; FastText: $p < 0.001$; ELMo: 0.001), except for PPMI.w2 ($p = 0.23$), PPMI.synf ($p = 1$), and BERT ($p = 0.39$). Conversely, the *AP* of FPs is significantly higher than the one of TNs (SGNS.w2: $p < 0.001$; SGNS.synf: $p < 0.001$; SGNS.synt: $p < 0.001$; FastText: $p < 0.001$), except for the largely underperforming PPMI models (PPMI.w2: $p = 1$; PPMI.synf: $p = 0.39$; PPMI.synf: $p = 0.38$), ELMo ($p = 1$), and marginally for GloVe ($p = 0.08$) and BERT ($p = 0.08$). This suggests that the semantic features of FPs tend to overlap with the top features of the positive class more than TNs.

The analysis of the semantic features of missclassified words can also provide interesting clues to explain why DSMs make errors in probing tasks. For instance, FastText does not classify *keen* as a sound emission verb (i.e., it is an FN for the VerbNet class 38). If we inspect its decoded vector we find COGNITION, SOCIAL, SELF, BENEFIT, and PLEASANT among its top features, likely referring to the abstract adjective *keen*, which is surely much more frequent in the PoS-ambiguous training data than the verb *keen* (to emit wailing sounds). On the other hand, PPMI.w2 wrongly classifies *judge* as a manner of speaking verb (i.e., it is an FP of the VerbNet class 37.3). This mistake can be explained by looking at its decoded vector whose top feature is SPEECH, which is probably due to the quite common usage of *judge* as a communication verb.

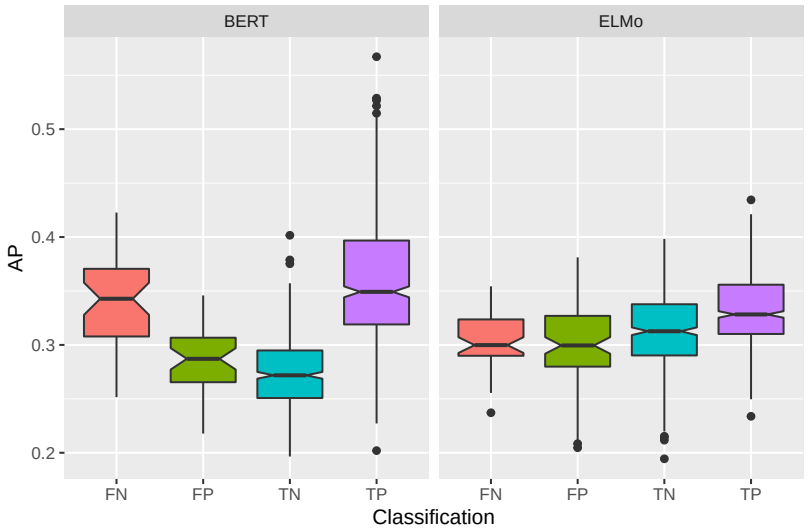


Figure 6
Average Precision (*AP*) boxplots of the Binder vectors of the test words with respect to the top-20 Binder features of each probing task. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) refer to the output of the classifiers trained on the original BERT and ELMo embeddings.

As illustrated in Tables 8 and 9, the variance of model accuracy across tasks is extremely high. For instance, the accuracy of FastText ranges from 1 in the Positive/Negative and Concrete/Abstract tasks, to 0.55 for the VerbNet say-37.7 class. In the standard use of probing tasks, the classifier accuracy is taken to be enough to draw conclusions about the way a certain piece of information is encoded by embeddings. Here, we go beyond this "black box" analysis and provide a more insightful interpretation of the different behavior of embeddings in semantic probing tasks. We argue that such explanation can come from the decoded Binder features, and that a model performance in a given task t depends on the way the words to be classified encode the top-n ranked features for t (i.e., $TopTaskFeats(t, n)$). For instance, consider the boxplots in Figure 7, which show the *AP* of the Binder vectors decoded from the FastText embeddings for the words belonging to the positive (1) and negative (0) classes in the test sets of the Positive/Negative and VerbNet say-37 probing tasks. FastText achieves a very high accuracy in the former task, and the *AP* distributions of the 1 and 0 words are clearly distinct, indicating that these two sets have different semantic features, and that the features of the 0 words have a very low overlap with the top task features. Conversely, the *AP* distributions of the 1 and 0 words for the say-37.7 task overlap to a great extent, suggesting that the two groups are not well separated in the semantic feature space. Our hypothesis is that the DSM accuracy in a probing task tends to be strongly correlated with the degree of separation between the semantic features decoded from the positive and negative items in the target class.

To verify this hypothesis, we take the sets of positive (W_1) and negative (W_0) test words of each task t and we compute the following measure:

$$AP_{diff}(t) = \overline{AP}(t, W_1) - \overline{AP}(t, W_0) \quad (4)$$

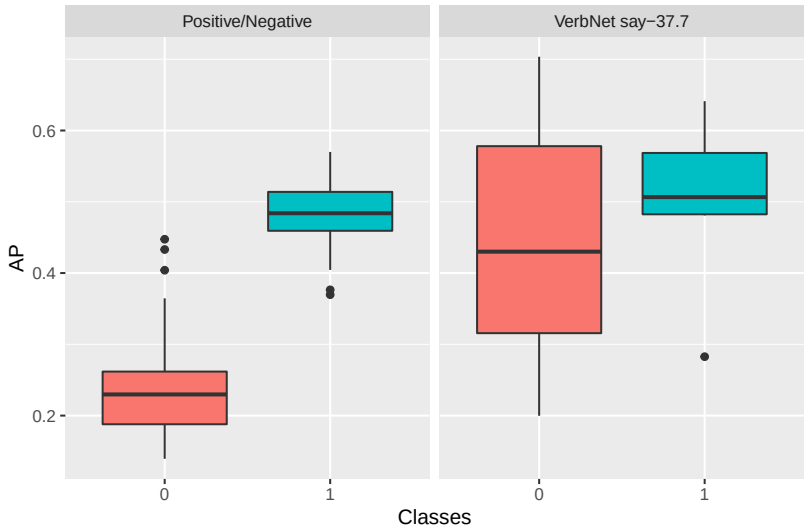


Figure 7
The boxplots show the Average Precision (*AP*) of the Binder vectors decoded from FastText embeddings for the words belonging to the positive (1) and negative (0) classes in the test sets of the Positive/Negative and VerbNet say-37 probing tasks.

where $\overline{AP}(t, W_1)$ and $\overline{AP}(t, W_0)$ are respectively the mean AP for W_1 and W_0. Therefore, $AP_{diff}(t)$ estimates the separability of the positive and negative words in the semantic feature space relevant for the task t. We expect that the higher the $AP_{diff}(t)$ of a model, the higher its performance in t. Table 11 shows that this prediction is borne out, at least for the best performing non-contextualized DSMs. The Spearman correlation between the model accuracy in the probing tasks and $AP_{diff}(t)$ is fairly high for all models, except for the PPMI ones. It is again suggestive that these are not only the worst-performing models in the probing tasks, but also the embeddings with a less satisfactory encoding of the Binder features. Table 12 illustrates that the correlation between $AP_{diff}(t)$ and task accuracy holds true for contextualized embeddings as well. For both BERT and ELMo, the AP_{Diff} and accuracy are greater for the Direct Object Animacy task than for the Causative/Inchoative alternation.

Table 11
Spearman correlation (ρ) between $AP_{diff}(t)$ and the classification accuracy for the non-contextualized embeddings models.

Model	ρ	**p-value**
PPMI.w2	0.29	0.15
PPMI.synf	0.43	0.03∗
PPMI.synt	0.23	0.26
GloVe	0.65	< 0.001∗
SGNS.w2	0.68	< 0.001∗
SGNS.synf	0.78	< 0.001∗
SGNS.synt	0.70	< 0.001∗
FastText	0.71	< 0.001∗

Table 12
Classification accuracy and $AP_{diff}(t)$ for the contextualized models.

Task	Model	Accuracy	$\mathbf{AP_{diff}}$
Direct object animacy	BERT	0.99	0.13
Direct object animacy	ELMo	0.96	0.04
Causative/Inchoative alternation	BERT	0.91	0.06
Causative/Inchoative alternation	ELMo	0.86	0.01

6. General Discussion and Conclusions

Word embeddings have become the most common semantic representation in NLP and AI. Despite their success in boosting the performance of applications, the way embeddings capture meaning still defies our full understanding. The challenge mainly depends on the apparent impossibility to interpret the specific semantic content of vector dimensions. Indeed, this is the essence of **distributed representations** like embeddings, in which information is spread among patterns of vector components (Hinton, McClelland, and Rumelhart 1986). Consequently, the content of embeddings is usually interpreted indirectly, by analyzing either the space of nearest neighbors, or their performance in tasks designed to "probe" a particular semantic aspect.

In this article, we have taken a different route, adopting a methodology inspired by the literature on neural decoding in cognitive neuroscience. The brain, too, represents semantic information in distributed patterns (Huth et al. 2016). We argue that the problem of interpreting the content of embeddings is similar to interpreting the semantic content of brain activity. Neurosemantic decoding aims at identifying the information encoded in the brain by learning a mapping from neural activations to semantic features. Analogously, we decode the content of word embeddings by mapping them onto interpretable semantic feature vectors. Featural representations are well-known in linguistics and cognitive science (Vigliocco and Vinson 2007), and provide a human-interpretable analysis of the components of lexical meaning. In particular, we rely on the ratings collected by Binder et al. (2016), whose feature set is motivated on a neurobiological basis. We have carried out the mapping of continuous embeddings onto discrete semantic features with a twofold aim: (i) identifying which semantic features are best encoded in word embeddings; and (ii) using the proposed featural representations to explain the performance of embeddings in semantic probing tasks.

Concerning the first goal, we have tested the embedding decoding method on several types of static and contextualized DSMs. All models achieve high correlations across words and features, with dependency-based DSMs having a slight edge over the others, consistently with the findings of Abnar et al. (2018). The features from abstract domains such as Cognition, Social, and Causal seem to be the ones that are better predicted by the models, which are purely relying on text-based information, while the prediction of spatial and temporal features is obviously more challenging. A further analysis reveals the salience of visual, motion, and audition features, supporting the hypothesis that language redundantly encodes several aspects of sensory-motor information (Louwerse 2008; Riordan and Jones 2011). In terms of word categories, the vectors are very good in predicting entities, whereas they struggle with physical and abstract properties. Moreover, it is interesting to observe that the new generation of

contextualized DSMs does not significantly differ from traditional ones for the type of semantic information they encode.

As for the second goal, we have applied our decoded feature representations to the widely popular probing task methodology, to gain insight on what pieces of semantic information are actually captured by probing classifiers. For our experiments, we tested the original embeddings on probing tasks designed to target affective valence, animacy, concreteness, and several verb classes derived from VerbNet for non-contextualized DSMs, and direct object animacy and causative/inchoative verb alternations for contextualized embeddings. If a binary classifier manages to identify whether a word belongs to a semantic class on the basis of its embedding, this is typically taken as indirect evidence that the embedding encodes the relevant piece of semantic information. In our work, instead of regarding probing tasks just as "black box" experiments, we use the decoded feature vectors to inspect the semantic dimensions learned by the classifiers. Moreover, we have set up a battery of tests to show how the decoded features can explain the embedding performances in the probing tasks. We have measured with AP the overlap between the top task features and the most important features of the test words belonging to the positive and negative classes. Our analyses reveal that:

- the words correctly classified in the positive class (i.e., TPs) share a large number of the top ranked features for that class, and, symmetrically, the words correctly classified in the negative class (i.e., TNs) have a significantly lower number of the top task features;
- words wrongly classified in the negative class (i.e., FNs) lack many of the top features characterizing the target class. Conversely, the features of words wrongly classified in the positive class (i.e., FPs) tend to overlap with the top task features more than TNs;
- the accuracy of a DSM in a probing task strongly correlates with the degree of separation between the semantic features decoded from its embeddings of the words in the positive and negative classes.

These results show that semantic feature decoding provides a simple and useful tool to explain the performance of word embeddings and to enhance the interpretability of probing tasks.

The methodology we have proposed paves the way for other types of analyses and applications. There are at least two prospective research extensions that we plan to pursue, respectively concerning selectional preferences and word sense disambiguation. Many recent approaches to the modeling of selectional preferences have given up on the idea of characterizing the semantic constraints of predicates in terms of discrete semantic types, focusing instead on measuring a continuous degree of predicate-argument compatibility, known as *thematic fit* (McRae and Matsuki 2009). DSMs have been extensively and successfully applied to address this issue, typically measuring the cosine between a target noun vector and the vectors of the most prototypically predicate arguments (Baroni and Lenci 2010; Erk, Padó, and Padó 2010; Lenci 2011; Sayeed, Greenberg, and Demberg 2016; Santus et al. 2017; Chersoni et al. 2019; Zhang, Ding, and Song 2019; Zhang et al. 2019; Chersoni et al. 2020; Pedinotti et al. 2021). This approach can be profitably paired with our decoding methodology to identify the most salient features associated with a predicate argument. For instance, we can expect that *listen* selects for direct objects in which Audition features are particularly salient. This way, distributional methods will be able not only to measure the gradient preference of

a predicate for a certain argument, but also to highlight the features that explain this preference, contributing to characterizing the semantic constraints of predicates.

As for word-sense disambiguation, models like ELMo and BERT provide contextualized embeddings that allow us to investigate word sense variation in context. Using contextualized vectors, it might be possible to investigate how meaning changes in contexts by inspecting the feature salience variation of different word tokens. For example, we expect features like SOUND and MUSIC to be more salient in the vector of *play* in the sentence *The violinist played the sonata*, rather than in the sentence *The team played soccer*. This could be extremely useful also in tasks such as metaphor and token-level idiom detection, where it is typically required to disambiguate expressions that might have a literal or a non-literal sense depending on the context of usage (King and Cook 2018; Rohanian et al. 2020).

Word embeddings and featural symbolic representations are often regarded as antithetic and possibly incompatible ways of representing semantic information, which pertain to very different approaches to the study of language and cognition. In this paper, we have shown that the distance between these two types of meaning representation is shorter than what appears prima facie. New bridges between symbolic and distributed lexical representations can be laid, and used to exploit their complementary strengths: *The gradience and robustness of the former and the human-interpretability of the latter*. An important contribution may come from collecting more extensive data about feature salience. The Binder data set is an important starting point, but human ratings about other types of semantic features and words might be easily collected with crowdsourcing methods.

In this work, we have mainly used feature-based representations as a heuristic tool to interpret embeddings. An interesting research question is whether decoded features from embeddings could actually have other applications too. For instance, semantic features provide a more abstract type of semantic representation that might be complementary to the fine-grained information captured by distributional embeddings. This suggests exploring new ways to integrate symbolic and vector models of meaning.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful feedback, and Yujie Qian for his support in setting up the experiments.

References

Abnar, Samira, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66, Salt Lake City, UT.

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR*, pages 1–13, Toulon.

Anderson, Andrew James, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev D. S. Raizada. 2016. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395. https://doi.org/10.1093/cercor/bhw240

Anderson, Andrew James, Edmund C. Lalor, Feng Lin, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Xixi Wang. 2018. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cerebral Cortex*, 29(6):2396–2411. https://doi.org/10.1093/cercor/bhy110

Athanasiou, Nikos, Elias Iosif, and Alexandros Potamianos. 2018. Neural

activation semantic models: Computational lexical semantic models of localized neural activations. In *Proceedings of COLING*, pages 2867–2878, Santa Fe, NM.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC*, 2010, pages 2200–2204, Valletta.

Bakarov, Amir. 2018. Can eye movement data be used as ground truth for word embeddings evaluation? In *Proceedings of the LREC Workshop on Linguistic and Neurocognitive Resources*, Miyazaki.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226. https://doi.org/10.1007/s10579-009-9081-4

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, MD.

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721. https://doi.org/10.1162/coli_a_00016

Beinborn, Lisa, Samira Abnar, and Rochelle Choenni. 2019. Robust evaluation of language-brain encoding experiments. *arXiv preprint arXiv:1904.02547*.

Binder, Jeffrey R., Lisa L. Conant, Colin J. Humphries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, and Rutvik H. Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4):130–174.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Boleda, Gemma and Katrin Erk. 2015. Distributional semantic features as semantic primitives - or not. In *Proceedings of Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches: Papers from the 2015 AAAI Spring Symposium*, pages 2–5, Stanford, CA.

Bommasani, Rishi, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of ACL*, pages 4758–4781, online.

Bradley, Margaret M. and Peter J. Lang. 2017. Affective Norms for English Words (ANEW). In *Technical Report C-3. UF Center for the Study of Emotion and Attention*, Gainesville, FL.

Buechel, Sven and Udo Hahn. 2018. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings of COLING*, pages 2892–2904, Santa Fe, NM.

Bulat, Luana, Stephen Clark, and Ekaterina Shutova. 2017a. Modelling metaphor with attribute-based semantics. In *Proceedings of EACL*, pages 523–528, Valencia.

Bulat, Luana, Stephen Clark, and Ekaterina Shutova. 2017b. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of EMNLP*, pages 1081–1091, Copenhagen.

Bulat, Luana, Douwe Kiela, and Stephen Christopher Clark. 2016. Vision and feature norms: Improving automatic feature norm learning through cross-modal maps. In *Proceedings of NAACL-HLT*, pages 579–588, San Diego, CA.

Bullinaria, John A. and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907. https://doi.org/10.3758/s13428-011-0183-8

Cardoso, Pedro Dias and Anindya Roy. 2016. Sentiment lexicon creation using continuous latent space and neural networks. In *Proceedings of the NAACL Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 37–42, San Diego, CA.

Carota, Francesca, Nikolaus Kriegeskorte, Hamed Nili, and Friedemann Pulvermüller. 2017. Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cerebral Cortex*, 27(1):294–309. https://doi.org/10.1093/cercor/bhw379

Chang, Kai min Kevin, Tom M. Mitchell, and Marcel Adam Just. 2011. Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *NeuroImage*, 56(2):716–727. https://doi.org/10.1016/j.neuroimage.2010.04.271

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Chersoni, Emmanuele, Ludovica Pannitto, Enrico Santus, Alessandro Lenci, and Chu-Ren Huang. 2020. Are word embeddings really a bad fit for the estimation of thematic fit? In *Proceedings of LREC*, pages 5708–5713, Marseille.

Chersoni, Emmanuele, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2019. A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, 25(4):483–502. https://doi.org/10.1017/S1351324919000214

Chronis, Gabriella and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of CoNLL 2020*, pages 227–244, online.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL*, pages 2126–2136, Melbourne.

De Choudhury, Munmum, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! Exploring human emotional states in social media. In *Proceedings of ICWSM*, pages 1–8, Dublin.

Derby, Steven, Paul Miller, and Barry Devereux. 2019. Feature2Vec: Distributional semantic modelling of human property knowledge. In *Proceedings of EMNLP*, pages 5853–5859, Hong Kong.

Devereux, Barry, Colin Kelly, and Anna Korhonen. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL Workshop on Computational Neurolinguistics*, pages 70–78, Los Angeles, CA.

Devereux, Barry J., Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4):1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, MN.

Djokic, Vesna, Jean Maillard, Luana Bulat, and Ekaterina Shutova. 2019. Modeling affirmative and negated action processing in the brain with lexical and compositional semantic models. In *Proceedings of ACL*, pages 5155–5165, Florence.

Erk, Katrin, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763. https://doi.org/10.1162/coli_a_00017

Esuli, Andrea and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, Genoa.

Ettinger, Allyson, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 134–139, Berlin, Germany.

Făgărăşan, Luana, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In *Proceedings of IWCS*, pages 52–57, London, UK.

Gauthier, Jon and Anna Ivanova. 2018. Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv:1806.00591*.

Glasgow, Kimberly, Matthew Roos, Amy Haufler, Mark Chevillet, and Michael Wolmetz. 2016. Evaluating semantic models with word-sentence relatedness. *arXiv preprint arXiv:1603.07253*.

Güçlü, Umut and Marcel A. J. van Gerven. 2015. Semantic vector space models predict neural responses to complex visual stimuli. *arXiv preprint arXiv:1510.04738*.

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, pages 4129–4138, Minneapolis, MN.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. https://doi.org/10.1162/COLI_a_00237

Hinton, Geoffrey E., James L. McClelland, and David E. Rumelhart. 1986. Distributed representations. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, MIT Press, Cambridge, MA, pages 77–109.

Hollenstein, Nora, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of CONLL*, pages 538–549, Hong Kong.

Howard, Jeremy and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL*, pages 328–339, Melbourne.

Huth, Alexander G., Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458. https://doi.org/10.1038/nature17637

Jackendoff, Ray. 1990. *Semantic Structures*, volume 18, The MIT Press, Cambridge, MA.

Jawahar, Ganesh, Benoît Sagot, Djamé Seddah, Samuel Unicomb, Gerardo Iñiguez, Márton Karsai, Yannick Léo, Márton Karsai, Carlos Sarraute, and Éric Fleury. 2019. What does BERT learn about the structure of language? In *Proceedings of ACL*, pages 3651–3657, Florence.

Kann, Katharina, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of SCIL*, pages 52–57, London, UK.

Kim, Najoung, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of *SEM*, pages 235–249, Minneapolis, MN.

King, Milton and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of ACL*, pages 345–350, Melbourne.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resource and Evaluation*, 42(1):21–40. https://doi.org/10.1007/s10579-007-9048-2

Klafka, Josef and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of ACL*, pages 4801–4811, online.

Kotlerman, Lili, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Journal of Natural Language Engineering*, 16(4):359. https://doi.org/10.1017/S1351324910000124

Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, Thomas K., Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ.

Lenci, Alessandro. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66, Portland, OR. https://doi.org/10.1111/tops.12335

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171.

Lenci, Alessandro, Gianluca E. Lebani, and Lucia C. Passaro. 2018. The emotions of abstract words: A distributional semantic analysis. *Topics in Cognitive Science*, 10(3):550–572.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308, Baltimore, MD.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word

embeddings. In *Transactions of the ACL*, 3:211–225.

Linzen, Tal, Grzegorz Chrupała, and Afra Alishahi. 2018. Introduction. In *Proceedings of EMNLP Workshop on BlackBoxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels.

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes. 2019. Introduction. In *Proceedings of ACL Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL*, pages 1073–1094, Minneapolis, MN.

Louwerse, Max M. 2008. Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15(4):838–844. https://doi.org/10.3758/PBR.15.4.838

Mandera, Paweł, Emmanuel Keuleers, and Marc Brysbaert. 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. https://doi.org/10.1016/j.jml.2016.04.001

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, Baltimore, MD.

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, Long Beach, CA.

McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559. https://doi.org/10.3758/BF03192726

McRae, Ken and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, 3(6):1417–1429. https://doi.org/10.1111/j.1749-818X.2009.00174.x

Mikolov, Tomas, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ.

Mitchell, Tom M., Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195. https://doi.org/10.1126/science.1152876

Murphy, Brian, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of *SEM*, pages 114–123, Montreal.

Murphy, Gregory. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Murphy, M. Lynne. 2010. *Lexical Meaning*. Cambridge University Press, Cambridge, UK.

Naselaris, Thomas, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. 2011. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nielsen, Finn Årup. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Palmer, Martha, Claire Bonial, and Jena D Hwang. 2017. VerbNet: Capturing English verb behavior, meaning and usage. *The Oxford Handbook of Cognitive Science*, pages 315–336.

Pedinotti, Paolo, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? Challenging transformers with generalized event knowledge. In *Proceedings of *SEM*, Online.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In

Proceedings of EMNLP, pages 1532–1543, Doha.

Pereira, Francisco, Matthew Botvinick, and Greg Detre. 2013. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence*, 194:240–252. https://doi.org/10.1038/s41467-018-03068-4

Pereira, Francisco, Greg Detre, and Matthew Botvinick. 2011. Generating text from functional brain images. *Frontiers in Human Neuroscience*, 5:72.

Pereira, Francisco, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963. https://doi.org/10.1016/j.neuron.2011.11.001

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, New Orleans, LA.

Poldrack, Russell A. 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692–697. https://doi.org/10.1016/j.neuron.2011.11.001

Pustejovsky, James and Olga Batiukova. 2019. *The Lexicon*. Cambridge University Press, Cambridge.

Recchia, Gabriel and Max M. Louwerse. 2015. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 68(8):1584–1598. https://doi.org/10.1080/17470218.2014.941296

Riordan, Brian and Michael N. Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Rohanian, Omid, Marek Rei, Shiva Taslimipoor, and Le Han Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of ACL*, pages 2890–2895, online.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20:33–53.

Santus, Enrico, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring thematic fit with distributional feature overlap. In *Proceedings of EMNLP*, pages 648–658, Copenhagen.

Sayeed, Asad, Clayton Greenberg, and Vera Demberg. 2016. Thematic fit evaluation: An aspect of selectional preferences. In *Proceedings of the ACL Workshop on Evaluating Vector-Space Representations for NLP*, pages 99–105, Berlin.

Schwartz, Dan and Tom Mitchell. 2019. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. In *Proceedings of NAACL*, pages 43–57, Minneapolis, MN.

Schwarzenberg, Robert, Lisa Raithel, and David Harbecke. 2019. Neural vector conceptualization for word vector space interpretation. In *Proceedings of the NAACL Workshop on Evaluating Vector Space Representations*, pages 1–7, Minneapolis, MN.

Sedinkina, Marina, Nikolas Breitkopf, and Hinrich Schütze. 2019. Automatic domain adaptation outperforms manual domain adaptation for predicting financial outcomes. In *Proceedings of ACL*, pages 346–359, Florence.

Şenel, Lütfi Kerem, Ihsan Utlu, Veysel Yücesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779. https://doi.org/10.1109/TASLP.2018.2837384

Shwartz, Vered and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the ACL*, 7:403–419. https://doi.org/10.1162/tacl_a_00277

Sikos, Jennifer and Sebastian Padó. 2019. Frame identification as categorization: exemplars vs prototypes in embeddingland. In *Proceedings of IWCS*, pages 295–306, Gothenburg.

Søgaard, Anders. 2016. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the ACL Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin.

Sun, Jingyuan, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of AAAI*, volume 33, pages 7047–7054, Honolulu, HI.

Tekiroglu, Serra Sinem, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of EMNLP*, pages 1511–1521, Doha.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of ICLR 2019*, pages 235–249, New Orleans, LA.

Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346. https://doi.org/10.1145/944012.944013

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188. https://doi.org/10.1613/jair.2934

Turton, Jacob, David Vinson, and Robert Smith. 2020. Extrapolating Binder style word embeddings to new words. In *Proceedings of the LREC Workshop on Linguistic and Neurocognitive Resources*, pages 1–8, Marseille.

Utsumi, Akira. 2018. A neurobiologically motivated analysis of distributional semantic models. In *Proceedings of CogSci*, pages 1145–1150, Madison, WI.

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

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA.

Vigliocco, Gabriella, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. 2009. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247. https://doi.org/10.1515/LANGCOG.2009.011

Vigliocco, Gabriella and David P. Vinson. 2007. Semantic representation. In Gareth Gaskell, editor, *The Oxford Handbook of Psycholinguistics*. Oxford University Press, Oxford, pages 195–215.

Vinson, David P. and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190. https://doi.org/10.3758/BRM.40.1.183

Vulić, Ivan, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835. https://doi.org/10.1162/COLI_a_00301

Vulić, Ivan, Edoardo M. Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of EMNLP*, pages 7222–7240, online.

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the ACL*, 7:625–641. https://doi.org/10.1162/tacl_a_00290

Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS ONE*, 9(11):e112575.

Wiedemann, Gregor, Steffen Remus, Awi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of KONVENS*, Erlangen.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*, Oxford University Press, Oxford.

Yaghoobzadeh, Yadollah, Katharina Kann, Timothy J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of ACL*, pages 5740–5753, Florence.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763, Vancouver.

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

Zhang, Hongming, Jiaxin Bai, Yan Song, Kun Xu, Changlong Yu, Yangqiu Song, Wilfred

Ng, and Dong Yu. 2019. Multiplex word embeddings for selectional preference acquisition. In *Proceedings of EMNLP*, pages 5247–5256, Hong Kong.

Zhang, Hongming, Hantian Ding, and Yangqiu Song. 2019. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of ACL*, pages 722–731, Florence.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, Santiago.