Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research

Giovanni Cassani[1], Federico Bianchi[2], Giuseppe Attanasio[3,4], Marco Marelli[5,6], and Fritz Günther[7]

[1]Tilburg University, Tilburg, The Netherlands

[2]Stanford University, Stanford, California, USA

[3]Bocconi University, Milan, Italy

[4]Instituto de Telecomunicações, Lisbon, Portugal

[5]University of Milano–Bicocca, Milan, Italy

[6]NeuroMI, Milan Center for Neuroscience, Milan, Italy

[7]Humboldt-Universität zu Berlin, Berlin, Germany

Author Note

Corresponding Author:

Giovanni Cassani

g.cassani@tilburguniversity.edu

Warandelaan 2, room D124

5037AB, Tilburg

The Netherlands

All authors declare that they have no conflict of interest. All materials, datasets, and analysis scripts for the studies reported here are available at https://osf.io/94ena/?view_only=4e29d179b7664f27b6b7b7ba6b75edf7, except for those that are available elsewhere (e.g., English Lexicon Project, Semantic Priming Project, MEN dataset, ...) and those that cannot be shared due to licensing constraints (Corpus of Contemporary American English). This study was not preregistered. This manuscript has been distributed in pre-print form on PsyArXiV and is available at this URL: https://osf.io/preprints/psyarxiv/b45ys. The content of this work has also been presented at the Virtual Psycholinguistic Forum of the Chinese University of Hong Kong, at the 57th Annual Meeting of the Society for Mathematical Psychology (MathPsych) and the 22nd International Conference on Cognitive Modeling (ICCM) and at the Highlights in Language Sciences (HILS) conference in Nijmegen..

Giovanni Cassani https://orcid.org/0000-0003-3917-3315

Marco Marelli https://orcid.org/0000-0001-5831-5441

Federico Bianchi https://orcid.org/0000-0003-0776-361X

Giuseppe Attanasio https://orcid.org/0000-0001-6945-3698

Fritz Günther https://orcid.org/0000-0002-9205-6786

Abstract

Computational models of semantic representations have long assumed and produced a single static representation for each word type, ignoring the influence of linguistic context on semantic representations. Recent Large Language Models (LLMs) introduced in Natural Language Processing, however, learn token-level contextualised representations, holding promise to study how semantic representations change in different contexts. In this study we probe type- and token-level representations learned using a prominent example of such models, Bidirectional Encoder Representations from Transformers (BERT), for their ability to i) explain semantic effects found for isolated words (semantic relatedness and similarity ratings, lexical decision, and semantic priming), but critically also to ii) exhibit systematic interactions between lexical semantics and context, and iii) explain meaning modulations in context. Across a wide range of empirical studies on each of these topics, we show that BERT representations satisfy two desiderata for psychologically valid semantic representations: i) they have a stable semantic core which allows people to interpret words in isolation and prevents words to be used arbitrarily and ii) they interact with sentence context in systematic ways, with representations shifting as a function of their semantic core and the context. This demonstrates that a single, comprehensive model which simultaneously learns abstract, type-level prototype representations as well as mechanisms of how these interact with context can explain both isolated word effects and context-dependent variations. Notably, these variations are not limited to discrete word senses, eschewing a strict dichotomy between exemplar and prototype models and re-framing traditional notions of polysemy in semantic memory.

Keywords: Distributional semantics; Large Language Models; Computational Psycholinguistics; Mental lexicon; Context-dependent semantic representations

Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research

As is illustrated by the very sentence you are now reading, or really any other piece of text, words usually do not appear in isolation. At least after a certain age, the way we use language to communicate with others more often than not consists of longer segments, in forms ranging from short phrases over sentences to discourse. That is, the individual words we use typically occur in the *context* of other words. This context-dependency is crucial when we consider the most central property of words: their meaning, which determines their communicative purpose. Homonyms are an especially obvious case: Consider the meaning of *ball* in *For the third time, the worried player swung but missed the ball.* versus *The lonely maiden had great hopes as she dressed to go to the ball.* (Till et al., 1988). But context-dependency of meaning is not a limited phenomenon restricted to the 7% of word types that are homonyms (i.e., words with different unrelated meanings); in fact, Rodd et al. (2002) show that 80% of word forms in the English vocabulary are polysemous (i.e., they have different related senses, such as *paper*). However, the notion of "polysemy" still assumes that we can capture word meanings in context with a finite set of discrete senses (Li & Joanisse, 2021; Rodd, 2020). While a useful tool in lexicography, this is arguably still an oversimplification: In fact, it can be argued that really no two instances of a given word have the exact same meaning, even when the same "sense" of the word is used (Haber & Poesio, 2024; Mickus et al., 2020): Consider *chair* in *There is a really nice chair in my living room/my dinner room/my office/the museum/the art gallery/the castle*, where each chair is different and comes with different affordances (e.g., whether one can for example sit on it or not) or features about materials, shape, time in which it was built, and more. In fact, there is meaning variation even when the word refers to the very same entity (Zwaan et al., 2002): The *tomatoes* in *The young tomatoes are growing in my garden/I pluck the ripe tomatoes/I chop the tomatoes on my cutting board/You can still identify some of the tomatoes in my boiling pasta sauce* are quite different, even when speaking about the same set of tomatoes. Thus, although words are generally used to refer to categories, every particular instance in which a word is used is also that – a unique instance, different from any other. It is hence in principle possible to ascribe a unique meaning and meaning representation to each *token*, rather than only to each *type* of a word or to discrete word senses (Apidianaki, 2023).

In line with these arguments, it has been repeatedly pointed out that the traditional research focus on static type-level representations is inadequate or at least insufficient, and word meaning representations need to be considered (also) at the token level (Ambridge, 2020; Connell & Lynott, 2014; Yee & Thompson-Schill, 2016). Indeed, after reviewing a wide range of empirical studies on various context effects, ranging from the effects of lifetime

experience with concepts to the effects of the immediate linguistic context and task demands, Yee and Thompson-Schill (2016) conclude that "rather than being static, conceptual representations are constantly changing and are inextricably linked to their contexts" (p. 1015).

In such a perspective, following up the classical distinction in concept research (Murphy, 2002a; Rosch, 1978; Rosch & Mervis, 1975; E. E. Smith & Medin, 1981), each token of a word can be identified as an individual *exemplar* of a category, with the word type making up the *prototype*, an abstracted, aggregate representation of that category (Ambridge, 2020; Jamieson et al., 2018). There is still an ongoing debate as to whether prototype- or exemplar-based models (or hybrids between them) are to be preferred (for recent contributions, see for example Ambridge, 2020; Battleday et al., 2020; Jamieson et al., 2022). In general terms, prototype models capitalize on the invariance in experience, which in the limit affords a single representation which subsumes several, possibly infinite, individual instances of a same unit, and aim to capture precisely what is invariant. On the other end of the spectrum, exemplar models capitalise on the contextual variability of experience, which in the limit prevents the possibility of having a single abstract representation that subsumes different instances of a supposed same unit. The higher fidelity with which experience can be encoded makes these models appealing. On the one hand, a single representation for a given unit is very compact, but may not scale well to new encounters with said unit. On the other hand, exemplar models would store any experience, even if some aspects of it could be abstracted away.

In order to even just engage in this debate, we need comparable implementations of both types of model in the first place. For some modalities, such implementations are available: For example, in the visual modality, modern deep learning models provide us with representations for both individual images (i.e., exemplars) and aggregate category-level representations (i.e., prototypes) (Battleday et al., 2020, 2021; Günther et al., 2023); similarly, for the acoustic modality, recent (deep) neural networks can process individual sound snippets and encode them, allowing a post-hoc aggregation to derive prototype representations. In research on phonology, the tension between prototype and exemplar models has been long researched (J. Bybee, 2002; Frisch, 2017; Kohl, 1993), with evidence that suggests a role of both prototype abstract representations as well as exemplars. Milin et al. (2020) shows a first attempt at modeling how abstract phonological units may emerge from exposure to speech sounds, comparing pure instance-based learning (stored exemplars) and error-driven learning, suggesting the second fares better.

However, when it comes to meaning and semantics, virtually all (computational) models of word meaning currently used in psychological science are implemented at the type

level and not the token level (cf. Jamieson et al., 2018; except for a few noteworthy studies, described in the section Contextualizing semantic representations below). Consequently, these models are not well-suited to capture context-dependency. That said, recent developments in artificial intelligence research and natural language processing (NLP) have brought forward a class of highly popular language models that are explicitly geared to represent context-dependent meanings for individual word tokens (Devlin et al., 2019a; Peters et al., 2018; for an overview, see Liu et al., 2020). These models appear to strike a possibly interesting balance between abstraction and storage: They do aim to find the invariance in experience, deriving compact representations of certain aspects of reality, but they also allow such abstract representations to be skewed by idiosyncratic aspects of the specific experience, allowing a richer representation that captures the variability of experience with higher fidelity. Adopting these models in cognitive science and psycholinguistics more specifically promises to fill in this crucial gap which has been identified – but remained unaddressed – for a long time.

To introduce these models, we first survey current computational models of word meaning (specifically, distributional semantic models), presenting their strengths that render them popular candidates for modelling psychological representations of word meaning and semantic memory, but also their clear weaknesses and limitations when it comes to taking into account context. After reviewing some earlier proposals for context-dependent meaning representations, we then present how modern language models developed in the field of NLP – such as the *Bidirectional Encoder Representations from Transformers* (BERT) model (Devlin et al., 2019a) on which we focus – can serve as a distributional semantic model that produces fully contextualized token-level meaning representations. The empirical part of this paper then thoroughly investigates whether the BERT model can serve as a psychological model of contextualized word meaning representations, by using it to predict a wide array of linguistic and psychological phenomena. Code and resources to generate the embeddings we used are available on a GitHub repository.[1]

Distributional models of semantic memory and word meaning representation

When it comes to models of semantic memory and word meaning representation, we have a considerable range of quantitative and computationally-implemented models available to us (M. N. Jones et al., 2015; Kumar, 2021). These include models that directly use human ratings on various meaning dimensions as quantitative meaning representations (E. M. Buchanan et al., 2019; Devereux et al., 2014; Lynott et al., 2020; McRae et al., 2005),

[1] https://github.com/MilaNLProc/psycho-embeddings

models that estimate underlying semantic representations from behavioral data (Benedek et al., 2017; De Deyne et al., 2016; Kenett et al., 2017), or models that estimate semantic representations as the learning result from a certain input representing our experience (Andrews et al., 2009; Bruni, Tran, et al., 2014; Landauer & Dumais, 1997; Mandera et al., 2017).

The most prominent representative of the latter are a class of models known as (depending on the scientific field) distributional semantic models (DSMs), vector space models (VSM) of meaning, or word embedding models (for overviews from a psychological perspective, see Günther et al., 2019; M. N. Jones et al., 2015; Lenci, 2018)[2]. These models start from the *distributional hypothesis* that words with similar meanings occur in similar contexts (Harris, 1954; Sahlgren, 2008) – that is, they have a similar *distribution over contexts*. Once an operational definition of *context* is applied – such as, for example, a document a word occurs in (Landauer & Dumais, 1997) or the words immediately surrounding a target word (Lund & Burgess, 1996; Mikolov, Chen, et al., 2013) – these distributions can be extracted from large corpora of natural language. Although, traditionally, such distributions were learned and represented via count-based strategies, the affirmed approach today is using *continuous vectors* (Baroni et al., 2014a). Using this type of representation – i.e., a long list of numbers – for distributions over contexts means distributional models represent word meanings as high-dimensional vectors spanning a semantic space. Crucially, distributional semantic models derive these word meaning representations purely from how words are used and occur in language.

For corpora that reasonably approximate human language experience, distributional semantic models serve as psychologically plausible learning models that derive semantic representations from this input (Hollis, 2017; Lenci, 2008; Mandera et al., 2017). The plausibility of the resulting representations has been confirmed in a wide range of empirical studies predicting human behavioral data from distributional vectors and the similarities between them (e.g., Baroni et al., 2014b; Günther et al., 2016b; Louwerse, 2011; Mandera et al., 2017; Nematzadeh et al., 2017; Pereira et al., 2016). Because of these different factors, distributional semantic models can currently be considered state-of-the-art models of human semantic memory and semantic representation, and they are consequently widely employed in empirical and theoretical studies (Günther et al., 2019; Kumar, 2021).

Of course, their wide adoption does not mean that these models are without problems and limitations, and criticism has come from many different directions (for a discussion, see Günther et al., 2019). One critical limitation is the fact that, once these models have been

[2] As in the literature, the terms *embedding* and *vector* will be used interchangeably in this article

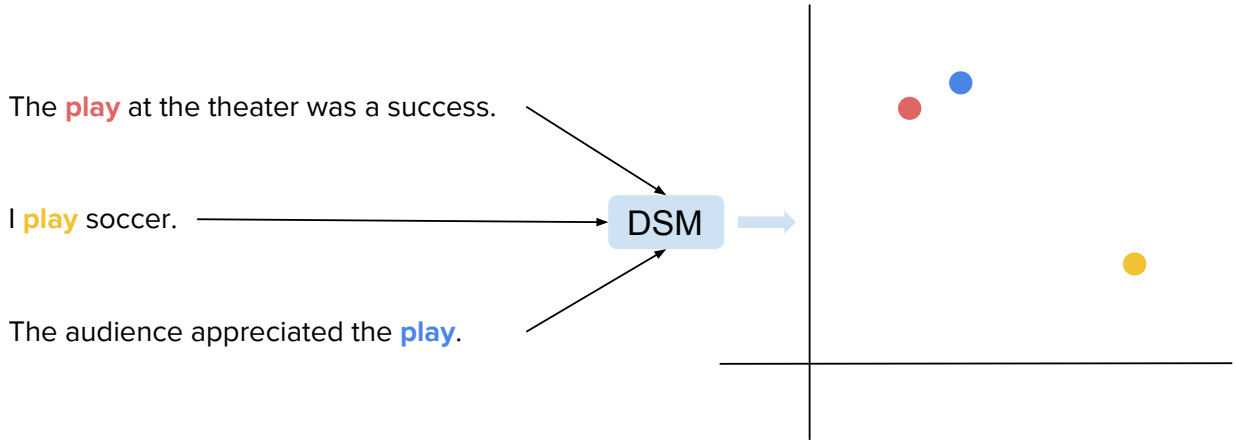


Figure 1

Depending on the context in which a word occurs, it is embedded in a geometrical space (here 2-dimensional for simplicity, but multidimensional in real applications) where representations are closer in space when context of occurrence is similar and further apart when context of occurrence differs. This typically applies to different types, but in more recent models like the one at the center of the paper this also applies to token-level representations.

trained on a given corpus, they represent each word meaning with a single static vector – which raises the question how this format can be justified for words with multiple different meanings and senses. Going beyond the phenomena of homonymy and polysemy, it has also been questioned how this fixed and static representation format can capture novel, context-specific uses of words, such as using a *newspaper* as an umbrella when it is raining (Glenberg & Robertson, 2000). However, it is important to emphasize here that, while these criticisms are usually directed at distributional semantic models, *all* models that provide static representations of word meaning at the type level – and that is, all current models of semantic memory described above – are faced with the very same problems.

As argued elsewhere, the fact that these models learn type-level representations does not imply that distributional vectors (or other static type-level representation models) only represent one single, discrete symbolic meaning (Günther et al., 2019; Günther & Marelli, 2022). Analyzing the semantic neighborhoods of distributional vectors often yields clusters that can be identified with different meanings or senses of a word (Camacho-Collados & Pilehvar, 2018; Heylen et al., 2015), and can be used to detect the "ambiguity status" (monosemous vs. polysemous vs. homonym) of words (Beekhuizen et al., 2021). In a recent study, Günther and Marelli, 2022 also showed that frequency distributions of different qualitative interpretations for compound words (a *wood brush* can be interpreted as a *brush FOR wood* or *brush MADE OF wood*) can be predicted from the dimensional values of their

distributional vectors, showing that this meaning variability is encoded in these vector representations.

However, while these studies show that different meanings are subsumed in static distributional vectors, these vectors are still fully abstracted: they simultaneously carry with them all these different meanings in a single representation, without naturally being able to tell them apart or to indicate which meaning is relevant at a given point in a given context. Consequently, the semantic similarity between the vectors for *ball* and *prom* is a fixed value, as is the semantic similarity between *ball* and *tennis*, even though these words hardly have any relation to the respective other meaning of *ball*. A context-sensitive model would instead tell us that *prom* is similar to *ball* in *The lonely maiden had great hopes as she dressed to go to the ball.* but not in *For the third time, the worried player swung but missed the ball.*, and vice versa for *tennis*.

Contextualizing semantic representations

Previous models of semantic representations

In the domain of cognitive science, some tentative approaches for obtaining context-sensitive word meaning representations have already been put forward. The earliest proposal of how regular, static type-level distributional vectors can be adjusted based on their context is Kintsch's Predication algorithm (Kintsch, 2001), originally proposed for metaphor representations (Kintsch, 2000). This algorithm is designed for predicate-argument structures such as IS.SHARK(LAWYER) (in *My lawyer is a shark*) or RUN(HORSE) (in *The horse is running*), and specifically to capture how the presence of the argument affects the meaning of the predicate (*run* has a different meaning in *The horse/the color/time is running*). In Kintsch's model, this context-sensitivity is realized by activating the m (a large number) nearest neighbors to the predicate, then selecting from those the k (a small number, $k << m$) nearest words to the argument, and then adding the vectors for these k words to the predicate vector. This has the effect that the contextualized predicate vector is nudged towards the argument vector. Variations of this method that follow the same underlying logic (activation and selection of neighbors) have later been proposed by Utsumi (2011) and Harati et al. (2021), also with a focus on metaphor representation (see Nick Reid and Katz, 2018 for an overview). However, this model is limited in that it is only designed to consider a single other word as context, and it is not clear if and how it can be applied recursively to consider more context. In addition, as remarked by Jamieson et al. (2018), such approaches are ad-hoc additions to an existing semantic model so that it can account for a specific behavior, rather than an inherent part of its design.

In a recent synthesis of research in the connectionist tradition (such as Armstrong and Plaut, 2008, 2016; Rodd et al., 2004), Rodd (2020) also proposes a model of meaning disambiguation. Here, the entry point to word meaning access is a static high-dimensional semantic vector tied to the word form (in this model, a binary vector indicating the presence or absence of semantic features). This vector is conceptualized as a "relatively unhelpful blend state that will likely contain inconsistent elements of meanings that correspond to different possible interpretations of the incoming word form" (Rodd, 2020, p. 414), very similar to the static vectors in distributional semantic vectors outlined above. This model further assumes that some features have positive connections and co-activate each other, while others have inhibitory, negative connections. Importantly, context cues can also reinforce some features while inhibiting others. For polysemous words, the simultaneous activation of incongruent features in the form-based vectors initially results in an unstable state, which can however be resolved by such context cues (for example, the presence of the word *player* will further activate the sports-related features of *ball* while inhibiting the festivity-related features). The different word senses of a word then emerge as *attractor basins*, stable states of co-activated features. However, even if the model assumed that they are learned from experience, in practice these high-dimensional feature vectors are typically hand-coded, resulting in small-scale implementations with considerable amounts of degrees of freedom available to the modeller (Hummel & Holyoak, 2003).

In another recent study, Jamieson et al. (2018) propose an instance theory of semantic memory for which "disambiguation of word meaning, even when the meaning is subordinate, falls naturally out of first principles" (p. 133). This model assumes that individual experiences are stored in episodic memory, and that abstracted representations such as word meanings emerge during retrieval from memory rather than as a product of learning (see also Ambridge, 2020; Jamieson et al., 2022; Johns et al., 2020). In the implementation of this model (see Appendix B for mathematical details), traces of all encountered language examples (e.g., sentences or documents) are stored in episodic memory as the sum of random vectors that are each uniquely associated to a given word type. When a word is then presented to the model as a probe, it activates these traces to a level based on their similarity to the unique vector associated to the probe, and then produces the semantic representation for the word as a weighted sum of these activated traces. By presenting multiple probes at the same time (a *joint probe* such as *break/car*, *break/glass*, or *break/news*, the model will then produce different representations for these joint probes, thus disambiguating the word meaning. However, these contextualized representations are only available for the entire word-plus-context expression, not the individual words in this expression themselves. It can therefore be argued that the model actually provides

representations for multi-word expressions, rather than truly disambiguated representations for individual token-level word meanings.

Li and Joanisse (2021) proposed an implementation grounded in Recurrent Neural Networks (RNNs) and implemented using a Long-Short Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997), which processes sequential information and allows previous steps to influence the representation derived for the current step. Therefore, previous words in a sentence influence the representation of the current one. The influence of context is continuous rather than discrete, as in attractor models. The proposed RNN model learns by predicting the next word, and develops representations that improve this prediction (effectively a probability distribution over all possible vocabulary entries) via exposure to naturalistic language input. Crucially, the semantic representation of a token is the outcome of a non-linear integration of the preceding context and the stored abstract semantic knowledge related to the word being processed. The model derived clustered representations of words in context which reflect human annotations of different word senses, and was more effective when words had different unrelated senses. The model was also shown to capture context dependent similarities (e.g., that people perceive movie admission as close to concert admission than guilty admission) and reading times in sentence processing, where words read faster were assigned a higher probability by the model.

All these models share the goal to incorporate the influence of context on semantic representations, following different routes that span exemplar-based models (Jamieson et al., 2018), variation over prototype models (Kintsch, 2000), attractor models (Rodd, 2020) and learning models (Li & Joanisse, 2021). In the next section, we describe the architecture of the contextualised model of meaning we evaluate in your study, which, we believe, holds promise to study the interplay between variance and invariance in semantic representations and indicates another possible account, with similarities to all previously mentioned models. At a very general level, BERT strikes what we hypothesize is a productive balance between learning the invariance in experience as well as a model of how context modulates representations in systematic ways.

Inherently contextualized language models

BERT (Devlin et al., 2019b) is a deep neural network that makes use of the Transformer architecture (Vaswani et al., 2017). In its core underlying assumptions and principles, BERT resembles other distributional semantic models (see also Lenci et al., 2022), which however – as we discussed before – only learn a static type-level representation. BERT, on the other hand, further learns to adapt an initial static representation to its current sentence context by exploiting information provided by co-occurring tokens in the same sentence. In this

section, we will describe how BERT works in general, and how it arrives at context-sensitive token-level representations.

Akin to multivariate linear regression models, neural networks consist of a set of input variables and a set of output variables – each variable being a "node" of the network – and learn the "regression" weights between those nodes that best predict output from input. Neural networks can have a hidden layer between input and output, which effectively functions as intermediate "output" for the input layer, and as intermediate "input" for the output layer. Such a hidden layer allows the neural network to also model non-linear relations between input and output. It is also possible to have multiple hidden layers, where each layer receives input from the previous one(s) and passes on its output as input to the next layer(s).

In language modelling, one highly prominent example of such a neural network with a hidden layer is *word2vec* (Mikolov, Sutskever, et al., 2013), which we will briefly explain to illustrate some core concepts that are relevant for BERT. The input layer as well as the output layer for *word2vec* contains one node for each word that appears (sufficiently often) in a language corpus. The hidden layer contains a fixed number of N nodes, for large corpora usually $N = 300$ or $N = 400$. The network is then trained to predict, as accurately as possible, which words (output) occur in the immediate surrounding of a given target (input) word, or vice versa.[3] During network training, this is done successively for each word token in the corpus as the target word. At the end of this training process, all weights between the nodes are fixed. This has one crucial implication: When setting one specific word in the input to 1 (and all others to 0), the fixed weights between input layer and hidden layer specify all values in the hidden layer. As this hidden layer is then used to predict which other words are how likely to occur in the immediate surrounding of the target, this hidden layer captures the "distributional history" of the target word as a N-dimensional vector. This vector is called a *distributional vector* or a *word embedding*, and as in other distributional semantic models it is taken as a word meaning representation (Mandera et al., 2017; Mikolov, Sutskever, et al., 2013). Such vectors can be obtained (or *extracted*) from the network for all words in the input layer.

Neural networks with more than one hidden layer are referred to as *deep neural networks* (see LeCun et al., 2015). The *transformer architecture* is a special type of deep neural network which has been popularized in the seminal paper by Vaswani et al. (2017), and includes several key components that we briefly outline here.

---

[3] To this end, the values of nodes for words that are present in the input or output are set to 1, and all other nodes are set to 0.

Given a sentence $S = \{w_1, w_2, \ldots, w_n\}$, BERT divides it into (sub-)lexical units $U = \{u_1, u_2, \ldots, u_m\}$. Unlike as discussed for *word2vec*, these units can but do not have to be full words; most of the time, they will be sub-lexical units which do not have to be actual morphemes (for example, *indoctrinate* might be divided into *ind*, *oct*, *ri*, and *nate*. This enables the model to deal with out-of-vocabulary words by relying on sub-lexical regularities in language: In order to compute the representation for a word, the representations of its sub-word tokens are simply averaged. For simplicity, we will still describe the model as if it was trained on *words* instead of these sub-lexical units in the remainder of this article. During training, one or more words are going to be *masked* and replaced by the special token [MASK], signaling the word to be predicted from its context (similar to *word2vec*, as described before). To each word in the sentence corresponds an embedding (i.e., a distributional vector) at layer 0, the most superficial layer in BERT's architectures, which is the type-level embedding for the word in the input sentence. The word is thus a pointer to a static semantic space, conceptually similar to the *word2vec* type-level embeddings discussed before. Hussain et al. (2023) provide an accessible tutorial on how to use models like BERT for behavioral research, as well as a detailed explanation of the main classes of such models, their learning principles, and different architectures.

Essentially, BERT is trained to recognize which words have been masked: given a sentence in input (such as, "the cat is on the table") one of the words is *masked* ("the cat is on the [MASK]") and BERT has to predict that the masked word is indeed *table*. The prediction of masked words is not the only learning objective by which BERT is trained, however. Next to this task, it is also trained to predict the next sentence. This second objective aims to train the model to develop representations that are sensitive to the coherence of broader linguistic contexts. Therefore, BERT's embeddings develop to allow the model to be sensitive to local and broader linguistic contexts. This process happens through a succession of layers (the exact number depends on specific implementations): at each successive layer, the representation integrates more and more information from the sentence context.

To make the model aware of sequential information, BERT also injects positional information through the use of an additional positional-specific embedding added to the static embedding of a word, which superficially (without taking into account the actual words in a sentence) distinguishes words at the first, second, ..., n-th position in the sentence. The positional vector is combined with the static embedding of a word and then these embeddings are passed through several so-called transformer encoder blocks (Vaswani et al.,

2017), a neural network that uses a technique called *attention*[4] to learn on which parts of the input sentence to focus, assigning more weights to those that matter. This device implements the notion that different words are more or less informative in delineating the context in which a target is used, and thus improve the prediction of the masked word. In practice, the attention mechanism generates an output representation for word i by computing a weighted average of the representations of all the other words in the sentence. These weights are in turn defined by the similarity between the different word representations in the embedded space. The parameters in each transformer encoder are crucially learned during training rather than being pre-specified: these weights encode the information the model has acquired over training about how context influences the representation of a masked word so that the right word-form can be predicted. Thus, BERT's context-sensitive representations for a word in a sentence are the vector sum of a static word embedding (i.e., a traditional distributional vector), a vector that blindly encodes the word's position in the sentence, and this weighted sum of the embeddings of other words in the sentence, taking into account how informative these words are for predicting the word in question. The output of the final layer is a series of embeddings for each word: continuous, numerical vectors which all exist in the same embedding space and can thus be compared also across layers. BERT thus enables us to extract word representations that are fully contextualized: when considering the model's output, there is no one-to-one correspondence between word type and word embedding (as in static DSMs) but rather a one-to-many mapping: depending on the linguistic context in which it is used, the same word can and will have different vector representations/embeddings. Therefore, BERT learns token-leven representations in addition to type-level word embeddings.

Figure 2 presents a high-level scheme of the process used to train BERT.

The BERT model (as well as the transformer architecture with which it is implemented) has been introduced to address some of the shortcomings of the LSTM network (Li & Joanisse, 2021), which exploits very similar computational mechanisms: Both models learn representations by predicting masked words (and contextually learn a probability distribution over all vocabulary items given a certain context), both models derive abstract, type level representations that are then warped by context, both models learn which contextual information to assign more weight to in any given context to make better predictions. However, transformer architectures have been shown to learn other linguistic phenomena (such as short- and long-distance syntactic dependencies; Hewitt and Manning, 2019) that LSTM models could not. We thus contend it is more fruitful to explore

[4] In spite of the name, attention mechanisms in neural networks are not isomorphic to human attention and they primarily serve as a computational tool to integrate context into embedding representations.

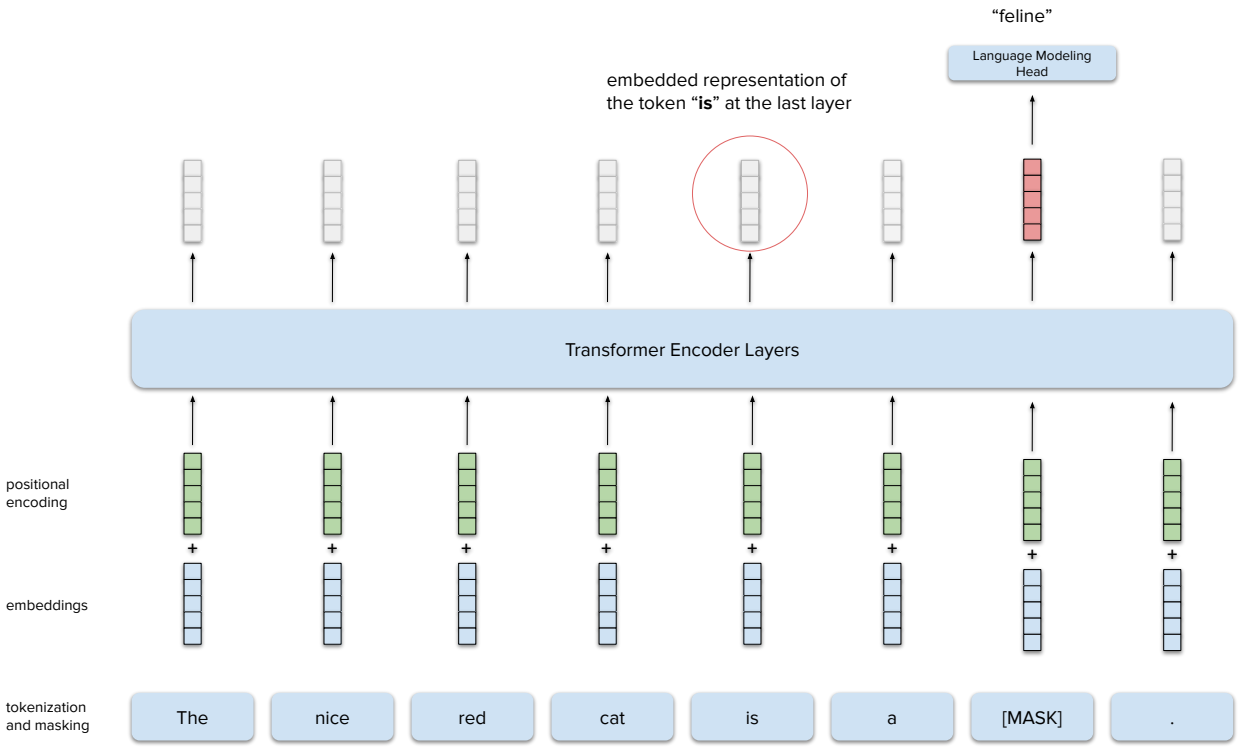


Figure 2

The sentence is first tokenized (each box at the bottom corresponds to one token) and used to retrieve the static embedding corresponding to the word type. This embedding is first combined with positional information and then traverses several transformer encoders which used neural attention to integrate linguistic context (in the form of words to the left and to the right of the masked token) and allow the model to correctly predict the masked token.

the capabilities of a model which has been reported to fit a wider range of linguistically relevant phenomena.

Contextualized Representations in Natural Language Processing

Contextualized representation models based on transformers have been incredibly successful in Natural Language Processing. One of the main grounds of evaluation, in recent years, has been provided by language understanding benchmarks like General Language Understanding Evaluation (GLUE; Wang et al., 2018) and SuperGLUE (Wang et al., 2019). These benchmarks contain multiple datasets that can be used to compare models' capability on many linguistic and semantic tasks, including coreference resolution, word sense disambiguation, question answering, sentiment analysis, linguistic acceptability, paraphrasing, and more. Compared to previous distributional models such as static word embeddings learned using each word's surrounding context (Mikolov, Chen, et al., 2013),

words in transformers models can access the *entire* context through attention (see Section Inherently contextualized language models). This paradigm shift allows, among other advantages, to build richer representations since every word is contributing, even if slightly, to the representation of every other word.

Several linguistic tasks have been shown to greatly benefit from this shift in representation learning. Consider, as an example, the sentence "*Yesterday, I visited grandma and I brought there a bunch of stuff. Also, I installed that Alexa device, as you asked. I have strong doubts that it will work but when you're ready, we can try to video-call her.*" where coreferents *grandma* and *her* are several words apart. A coreference resolution model based on word embeddings and LSTMs would struggle to learn such a long-range dependency. Conversely, BERT, loosely speaking, pays higher attention to *grandma* when contextualizing *her*, and vice-versa. Semantic tasks such as sentiment classification also benefit from contextualized representations. Take the sentence "*The movie has a great start, but it ends so badly it was not worth the ticket.*": the sentiment conveyed is different in different parts of the context, and contextualized representations are needed to learn from both. These models have thus provided a noticeable improvement to a number of tasks where long-distance and context dependencies are of primary importance.

BERT and following transformer models (e.g., Decoding-enhanced BERT (DeBERTa; He et al., 2021), Pathways Language Model (PaLM; Chowdhery et al., 2022)), have aced language understanding benchmarks, outperforming standard representations such as word embeddings and even surpassing the level of non-expert humans in some of the proposed tasks (Wang et al., 2019).

Adaptability has been one of the greatest success factors for contextualized models. Once models are pre-trained, i.e., they have learned how to create *meaningful* contextualized word representations, they can by fine-tuned (i.e., adapted) to solve specific tasks with minimal data and computational resources. As such, contextualized models have found widespread adoption in many fields, including computational social science (Card et al., 2022), medicine and healthcare (Rasmy et al., 2021; Roy & Pan, 2021), law and legal services (Chalkidis et al., 2020), and finance (Shah et al., 2022; Yang et al., 2020). For reference, the BERT pre-trained model was downloaded from the HuggingFace Hub[5] about 50 million times only in May 2023, and LEGAL-BERT, its specialized version for legal text understanding, 13 thousand times[6]. Such widespread adoption motivates our focus on bi-directional contextualized word representation models based on transformers, among

[5] https://huggingface.co/models

[6] https://huggingface.co/nlpaueb/legal-bert-base-uncased

which we choose BERT as the primary representative given its promising successes in modeling certain aspects of semantic processing, which we review in the next section.

Previous empirical studies on contextualised embeddings in psycholinguistics

Next to their astonishing empirical success across several NLP tasks, transformer-based models, of which BERT is one prominent instance, have already been extensively researched in psycholinguistics (Cevoli, Watkins, & Rastle, 2022; Ettinger, 2020; Rogers et al., 2020; Trott & Bergen, 2023; Wilcox et al., 2021). This body of research showed how this class of models improves over previous implementations on several tasks (although they still fail to capture some phenomena; Ettinger, 2020). However, so far they have been mostly used to investigate syntactic abilities and to operationalise words' surprisal in context (Cevoli, Watkins, & Rastle, 2022; Hao et al., 2020) to model reading times. A handful of studies, however, further investigated BERT's semantic representations and their ability to capture the organization of the mental lexicon (for a discussion of BERT in semantic memory modelling, see Kumar, 2021). We review them below, highlighting how our study connects with their design and insights.

Cevoli, Watkins, Gao, et al. (2022) started investigating contextualised semantic representations as a model to deal with semantic ambiguity and how ambiguous words are represented and processed, linking such models to the theory proposed by Rodd (2020). This study aims to test whether the level of ambiguity of a word-form (polysemous, homonymous, unambiguous) can be recovered from contextualised embeddings in an unsupervised way, offering promising, albeit limited, evidence that word forms with different ambiguity profiles had systematically different neighboring structures. This study, however, did not investigate contextualisation across the full network, but focused on the output representations of these models. We contend that an analysis of how representations are morphed by context can provide insight as to how contextualisation works and whether it aligns with psychological theories of meaning. Finally, this study did not yet compare contextualised and static representations with each other, nor attempted to validate whether the same model architecture yields psychologically plausible representations of words in isolation *and* in context, which we do in our work.

Trott and Bergen (2023) further used contextualised representations extracted using recent contextualised embedding models to study lexical meanings and how they interact with context. Using polysemous words and short, disambiguating contexts they study the degree to which word representations are categorical or continuous, coming to the conclusion that human behavior reflects both. Both sense boundaries extracted from curated resources as well as cosine distances derived from word vectors predicted reaction times in a primed

sensibility judgment task, in which participants had to determine which sentences made sense and which did not. Processing difficulty, indexed by reaction times in the sensibility rating, increased when participants saw an ambiguous word form as a prime and target used in different senses as well as when prime and target had a larger cosine distance in the reference, context-dependent embedding space. The authors take their evidence to counter frameworks positing the existence of a mental dictionary with separate entries for different senses (since continuous measures of semantic similarity also predict processing costs) as well as frameworks advocating for purely continuous meaning representations grounded in exemplar theories (because sense boundaries account for additional variance over continuous semantic similarity). Instead, they advocate for a hybrid model where both discrete senses and continuous relations influence processing and representations. However, this study only focuses on comparing contextualised representations, without considering the relation between abstract, type-level representations and context-dependent, token-level ones, a gap we address in our study.

Not specifically related to lexical meaning in context but more broadly to conceptual organization, Bhatia and Richie (2022) use BERT as a foundation model and fine-tuned it on human data to model a wide array of conceptual knowledge tasks, focusing on out of sample generalisation. Fine-tuning refers to the practice of further training a model, such as BERT, on a new dataset and possibly task to improve its underlying representations for a specific task or set of tasks, and has shown remarkable success in both NLP and cognitive science. More specifically, Bhatia and Richie (2022) use sentences generated from large-scale dataset of semantic feature and category norms to make the model better apt at using rich conceptual knowledge in a way that approximates representations and processes involved in simple semantic judgments involving thousands of naturalistic concepts and features. Their analysis encompasses '25 experiments involving response times in semantic verification tasks, typicality ratings for concepts and categories, the distribution of features across concepts, and similarity ratings between pairs of concepts' (p. 2). The authors further contrasted fine-tuned models with static word embeddings as well as the type-level embeddings derived by BERT, highlighting how fine-tuning on concept-feature relations improves the ability of the model to reproduce most semantic effects. Our work has several points of contact with this study, in that it focuses on semantic representations and uses transformer-based architectures. However, it also diverges in critical aspects: first and foremost, we aim to study representations of words in rich sentential contexts and especially how the interplay between lexical semantics and context morphs these representations. Moreover, while fine-tuning typically improves performance, we focus on the pre-trained model to investigate

whether the general-level learning mechanism employed by the BERT model already yields semantic representations that capture critical aspects of semantic processing.

Finally, preliminary evidence on the use of determiners in English (Madabushi et al., 2022) and the emergence of phones (Milin et al., 2020) suggests that learning models which rely on contextualised representations, of the kind we consider in this work, show a high degree of success. They seem to capture a certain degree of abstractions in how English determiners are used and show that phone-like units emerge from speech items. The cited works compared learning models with implementations of exemplar-based models (J. L. Bybee, 2013; Daelemans & Van Den Bosch, 2010), showing a better fit with respect to language data for the former. While some implementations of such learning models still rely on a prototype-based account, they allow to articulate the tension between a pure prototype-based and a fully exemplar-based account (Ambridge, 2020).

To sum up, like static, type-level DSMs, BERT aims to capitalize on the invariance in language experience, implementing the crude assumption that if the same lexical unit is used, it is to refer to the same unit in the world for the same purpose. However, it edges this assumption by relying on (linguistic) context to warp the representation tied to a certain lexical unit to be better able to predict co-occurring words. The model does not implement attractors explicitly nor does it posit their necessity for the model to converge into stable states. Since the model learns representations by predicting words in the sentence, it develops representations (of lexical items and context influences alike) that facilitate such prediction. The model also eschews storing every experience and rather derives contextualised representations on the fly by integrating abstract representations with the influence of context, both learned via exposure to naturalistic sentences. The model represents all words in the same way: polysemy is not hard-coded in the model, but is rather an emergent property (Cevoli, Watkins, Gao, et al., 2022), with each token representation occupying a portion of semantic space. If the token representations of a same word type are close together, the word type will have a rather unambiguous interpretation. The more the token representations of a same word type spread in semantic space, the more the word type is predicted to be ambiguous if presented outside of context (Cevoli, Watkins, Gao, et al., 2022).

Outlook for the present study

The aim of the present study is to establish whether word meaning representations obtained from the BERT model are psychologically plausible word meaning representations at both the token- and type-level. To this end, these representations must meet the following requirements:

1. The word meaning representations need to have a *semantic core*. We contend that

   (a) words should not be used entirely arbitrarily and thus not be arbitrarily replaceable in sentence contexts. The meanings of *bathtub* and *lion* in *The [bathtub/lion] hunts down the gazelle.* should not be identical, and bear at least some semblance to their meaning in isolation. At the same time, they should not be identical to their semantic core, as the context necessarily exerts an influence (when reading of bathtubs chasing gazelles, one might imagine a cartoonish bathtub with legs or a gazelle in the form of a plush toy to accommodate for the surprising situation)

   (b) words need yield a sensible meaning representation when presented in isolation. An extensive body of research has demonstrated semantic effects for words presented in isolation (e.g., L. Buchanan et al., 2001; Hino et al., 2002; Kuperman et al., 2014; Lachmair et al., 2011; Marelli et al., 2015; Pexman et al., 2008; Shaoul and Westbury, 2010), and any psychologically valid model of meaning is expected to account for this

2. The word meaning representations should be sensitive to the context the word appears in. That is, the representation for *ball* should be different in *For the third time, the worried player swung but missed the ball.* (more similar to *tennis*) versus *The lonely maiden had great hopes as she dressed to go to the ball.* (more similar to *prom*). However, meaning modulations should not just reflect different word senses. As exemplified in the *tomato* example at the beginning of this paper, context modulates meaning representations in systematic ways even when a word type is used in the same overarching sense in different contexts. We should thus further observe systematic shifts in meaning representations also for different tokens of the same word sense in different contexts.

We examine empirically whether BERT meets these criteria across a series of computational and statistical analyses of behavioral data, arranged in three sets of studies: (1) *single-word studies*, in which we examine isolated words without context and test whether type-level representations extracted from BERT account for variance in behavioral responses; (2) *replacement studies*, in which we replace a word in sentence contexts with other words (considering words with different degrees of semantic relatedness) and study how representations of targets and replacements shift based on context; and (3) *context variation studies*, in which we evaluate the same word in different contexts and investigate how the token-level representation changes with respect to both token- and type-level representations

as a function of context, considering word types used in different contexts to express a different sense or the same sense. Taken together, these studies allow us to examine how the context in which a word is presented on the one hand (requirement 2) and its identity on the other hand (requirement 1) influence its semantic representation, and how these semantic representations capture behavioral patterns.

This paper thus aims to empirically validate the psycholinguistic suitability of a model whose design allows a possible integration of type-level (prototype) and token-level (exemplar) representations. We move from a number of studies, reviewed in previous sections, showing that BERT succeeds at a number of tasks requiring contextualisation and aim to establish that this model can be fruitfully used to investigate aspects of human semantic processing by considering a series of related yet unaddressed questions. These cover how representations derived at different levels of the model fare in accounting for both explicit and implicit semantic effects, using ratings, reaction times, and free associations as behavioral measures. The design features we have highlighted above – chiefly, that the model learns a static, context-invariant representation for each lexical unit in its vocabulary as well as how to adapt it based on co-occurring lexical items through the neural attention mechanism – motivate the choice of such a model for this evaluation, that ultimately purports to study the validity of the underlying lexical representations from a psycholinguistic point of view. Furthermore, we explore how type- and token-level representations interact, which is especially interesting in the light of polysemy across and *within* sense boundaries, where both meaning stability and meaning modulations are key. Our goal is thus not to investigate whether contextualisation allows BERT to perform the way it does, for which ample evidence is available in the literature. Our empirical analysis instead targets the fundamental level of representations in a cognitive perspective and tackles questions such as whether word embeddings behave in a way that is useful for modeling semantic representations in a variety of tasks relevant for psycholinguistics.

All materials, datasets, and analysis scripts for the studies reported here are available at https://osf.io/94ena/?view_only=4e29d179b7664f27b6b7b7ba6b75edf7, except for those that are available elsewhere (e.g., English Lexicon Project, Semantic Priming Project, ...) and those that cannot be shared due to licensing constraints (Corpus of Contemporary American English). This study was not pre-registered. For most of the analyses we relied on existing datasets and our decisions on the sample size were driven by availability. Whenever this consideration does not apply, we specify how we addressed the issue of statistical power.

Token and Type representations

As we have mentioned in the previous section, transformer-based models learn a static representation which is then contextualised exploiting information provided by co-occurring tokens, information that helps the model predict which word was masked. In our simulations, we are interested in exploring the relation between type- and token-level representations learned by BERT. However, considering the initial, fully static embedding that BERT learns for a word is not the only way to obtain prototype embeddings given the representations the model yields. Following common practice in computer vision models (Günther et al., 2023), where representations of specific images depicting a concept are combined to obtain a prototype representation of the concept, vectors of specific word tokens in context can be averaged to obtain an abstract representation at the type level of the same word (Bommasani et al., 2020; Lenci et al., 2022). Unlike fully static embeddings, these representations combine context-dependent, token-level representations and have been often used as a proxy for type-level representations in NLP studies (Buijtelaar & Pezzelle, 2023; Lenci et al., 2022). The intuition is that abstractions at different layers capture different linguistic aspects, with more superficial layers focusing on form-related patterns, mid-layers capturing structural and semantic aspects of tokens, and deeper layers focusing on discourse- and sentence-level patterns (Rogers et al., 2020). By deriving type-level representations via averaging of token-level representations at different layers, we can thus investigate whether these representations are different than the fully static embedding BERT learns for each word and whether this bears relevance to psycholinguistic theories.

Next to averaging token-level representations and considering the initial, fully static representation on which contextualisation happens one can also derive type-level representations by presenting the model with the word in isolation, without any context, or in sentences which carry little to no contextual information, e.g. *This is a X* or *They X*, with X being replaced by any word form (Bommasani et al., 2020; Buijtelaar & Pezzelle, 2023). The assumption is that the absence of informative co-occurring words lets the model yield the representation of the word itself. Presenting the word in isolation can be used to simulate how the model would interpret isolated stimuli, replicating the experimental paradigms in many word-level psycholinguistic tasks. By feeding an isolated word and monitoring how its representation changes through BERT's layers we can study whether the model can also account for isolated word effects in a stringent design.

To sum up, in all simulations described in this paper, we work with the following underlying representations[7]:

- *contextualised representations:* these are token-level representations obtained by feeding a target word to the BERT embedder, and extracting its corresponding embedding. We have one embedding for each combination of word and context.
- *average prototypes:* these are type-level representations obtained by averaging 50 contextualised representations derived from as many sentences featuring the a given word as sampled from the Corpus of American English (CoCA, (Davies, 2009)) to obtain a single representation for each word type, akin to static embeddings. We thus have one such embedding for each target word, which collapses 50 random individual usages of the word type.
- *non-contextualised prototypes:* we fed each word for which we want to derive a representation to the BERT embedder in isolation, deriving a representation at each layer and obtaining a type-level representation directly by feeding BERT a word outside of any specific context. We thus have one such embedding for each target word.

For each of these token- and type-level representations, we obtained 14 embeddings:

- *static embedding:* this is the embedding at layer 0, before taking into account the positional encoding vector. To clarify, at this layer the token-level representation of the word *bank* in the sentences *Yesterday I went to the river bank to fish.* and *The bank closed its offices downtown and relocated.* will be exactly identical. This representation captures the lexical semantics BERT learns by being exposed to multiple sentences containing the same surface form (similar to a traditional static DSM). In all plots, this layer is indicated with the index -1. Moreover, this vector will be the same for contextualised, averaged prototype, and non-contextualised prototype representations since it is not affected by context in any way. This vector thus offers a consistent starting point to evaluate how context, even if only capturing the position of the token in the sentence, influences representations through the layers.
- *layer 0:* this is the embedding at layer 0, after including the positional vector. At this layer, representations for a same surface form will be identical only if the target word occurs at the same index in the sentence, so the representations for *bank* in the two

[7] If a word form is split in multiple tokens (e.g., *capybara* is tokenized as *cap*, *#y* and *#bara*), we average the embeddings of the tokens first.

example sentences above would differ as the target is the sixth token in the first sentence and the second token in the second sentence. This representation integrates a first, weak form of context.

- *layers 1-12:* each vector representation captures the token representation at different layers of the model, with layer 1 being superficial and 12 being the deepest layer in the architecture. Representations at deeper layers integrate more contextual information. We consider all layers to analyse to what extent the integration of sentence context affects representations across tasks, from those that depend on context more to those that only deal with isolated words.

We start our assessment of BERT's lexical representations by investigating whether prototype representations account for isolated word effects, considering three tasks: semantic relatedness and similarity norms (Bruni, Tran, et al., 2014; Hill et al., 2015), prime-target semantic relatedness effects in primed lexical decision data (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, Tse, et al., 2013; Koriat, 1981), and neighborhood effects in lexical decision data (Yap et al., 2011).

Isolated word effects

In this first suite of studies, we investigate the validity of BERT as a *type-level* model of semantic memory (probing requirement 1b outlined above that words should have a sensible meaning in isolation, i.e., when they are not presented in a sentence context). To this end, we test the model representations against several large-scale behavioral datasets in which words are presented in isolation, and that all play prominent and influential roles in testing models of human semantic memory: Explicit word similarity and relatedness ratings, lexical decision times in a semantic priming paradigm, and simple lexical decision times on single words. In all studies, we compare the performance of BERT type-level representations to a state-of-the-art model of semantic memory (Kumar, 2021; Mandera et al., 2017), the static distributional model *wordvec* (Mikolov, Sutskever, et al., 2013).

Experiment 1: Semantic relatedness and similarity norms

First, we focus on explicit off-line judgments, where participants are presented with word pairs and have to indicate their semantic similarity/relatedness. These ratings are often treated as the gold standard for semantic similarity scores derived from computational models (Baroni et al., 2014b; Lenci et al., 2022; Pereira et al., 2016).[8] Thus, models that

[8] However, from a psychological perspective, we advocate for a more cautious view, in which these judgments are seen as a type of behavior that, while heavily influenced by semantic similarity, is also susceptible to

better predict these judgments are taken to be better models of word meanings and lexical semantics.

In this first experiment, we replicate previous works and analyze the correlation between cosine similarities extracted from prototype embeddings obtained from BERT and semantic similarity/relatedness norms obtained by surveying human participants. Semantic similarity refers to paradigmatic relations between words, such that two words are semantically similar if they can be replaced with one another in context (Sahlgren, 2008). In this sense, similarity captures a rather narrow and specific semantic relation. Semantic relatedness, on the contrary, tackles syntagmatic relations: therefore, for two words to be semantically related, they have to co-occur together without the extra constraint that one can replace the other (Sahlgren, 2008): *coffee* and *drink*, for example, have high semantic relatedness but low semantic similarity. Semantic similarity and relatedness – also often referred to as a *semantic* versus *associative* relationships (M. N. Jones et al., 2006; Lucas, 2000) – are crucial to psychological theories of meaning (Kumar, 2021), and as such constitute an essential first step in our study.

Our analysis is closely related to previous studies (Lenci et al., 2022; Vulić et al., 2020), with some notable differences. First, we evaluate each layer separately rather than aggregations of the first layers, last layers, or only looking at the last layer. Moreover, we also crucially investigate to what extent non-contextualised prototypes capture semantic relations over layers. Our focus here is precisely on the comparison between averaged and non-contextualised prototypes, to establish which of the two approaches is better able to capture behavioral data and thus constitutes a better candidate for modelling human semantic representations. We do not have specific predictions about how the correlation should change over layers, but in order for BERT prototype embeddings to be usable to study semantic memory, the correlation should be generally high and robust.

Data

To analyse semantic similarity norms, we used the SimLex-999 dataset (Hill et al., 2015), computing the cosine similarity between each available pair of words (N = 990: For 9 word pairs, we did not find enough sentences in the CoCA containing either word, making it impossible to derive average prototypes. We thus opted to drop these pairs from the analysis.). SimLex-999 provides semantic similarity scores for 999 word pairs with different degrees of abstractness/concreteness and covering adjectives, nouns, and verbs. Word pairs were judged in isolation on a Likert scale and raters were instructed to rate pairs according

---

other influences. For these, and other reasons we articulate below, we also present an evaluation of BERT representations using other psycholinguistic tasks.

to how similar the two words are: the pair exhibiting the highest similarity is *vanish* - *disappear*, with a score of 9.80, whereas the pair with the lowest similarity is *new* - *ancient*, with a similarity of 0.23 (same value as the word pair *shrink* - *grow*). Inter-annotator agreement was 0.67, in line with similar benchmarks, and higher for abstract pairs and for adjective pairs. This dataset has been extensively used to benchmark models of meaning on their ability to encode semantic similarity.

To analyse semantic relatedness, we used the MEN dataset (Bruni, Tran, et al., 2014), another popular benchmark for computational models of meaning in NLP and cognitive science consisting of 3,000 word pairs. Unlike SimLex, MEN focuses on semantic relatedness, but the structure of the dataset is similar, with word pairs rated by native English speakers, this time on how related they are. Rather than through a Likert scale, ratings in the MEN dataset were collected using a comparative approach: raters were presented with two pairs and asked to indicate the one featuring the two words with the highest relatedness (e.g., given the pairs *car* - *automobile* and *giraffe* - *harbor*, participants most likely chose the former as more related). Word pairs are thus scored based on how many times they are picked as the most related pair in a trial. Each pair appeared in 50 trials, randomly matched to other pairs. The most related pairs were *car* - *automobile* and *sun* - *sunlight*, which were always chosen as the most related pair in a trial. The least related pair was *bakery* - *zebra* which was never chosen as the most related pair in a trial. We considered 2,973 word pairs, since for the remaining 27 pairs, one of the words did not occur in a sufficient number of sentences in the CoCA to derive the corresponding average prototype. We thus opted to drop these word pairs from the analysis.

Methods

In line with the approach used by the creators of both datasets to use them as a benchmarks for models of semantics (Bruni, Tran, et al., 2014; Hill et al., 2015), we extracted cosine similarities from computational models for the same word pairs appearing in the datasets and computed the Spearman rank correlation between the cosine similarity and the average semantic similarity/relatedness rating of each word pair. The higher the correlation between model-based cosine similarities and participants averaged ratings of semantic similarity/relatedness, the better a model's representations can be taken to encode the construct of interest.

As a comparison point for these various BERT embeddings, we used cosine similarities extracted from a reference*word2vec* space (Mandera et al., 2017) which has been previously validated on psycholinguistic tasks. The model is trained on a concatenation of

the UkWaC corpus and the SUBTLEX corpus, with 300-dimensional word vectors and a context window of 5 words to the left and right of the target word.

Results

Figure 3, upper panel, shows the correlations with SimLex-999 semantic similarity ratings the two prototyping approaches (averaging and non-contextualised), with the *word2vec* space as a baseline model. We see that the reference *word2vec* space (solid, light-blue line) shows a rather low correlation at 0.3. Average prototypes (dashed, dark-blue line) work rather well, with the correlation improving slightly over layers and remaining above 0.5, suggesting that averaged prototypes capture explicit similarity ratings moderately well. The picture is mixed for non-contextualised prototypes (long-dashed, blue line): at superficial layers, the correlation is only slightly lower than for average prototypes, but it declines towards deeper layers. This pattern suggests that when no context is available, BERT's deeper layers, trained to integrate context, are at a loss and the resulting embedding becomes unreliable. We further show that the correlation across cosine similarities between non-contextualised and averaged prototypes (dot-dashed, dark-red line) declines steadily from superficial to deep layers. Averaging correlates more strongly with the reference *word2vec* model (two-dashed, light-red line) than non-contextualised embeddings (dotted, red line), and this correlation follows closely that between SimLex-999 ratings and non-contextualised embeddings.

Figure 3, lower panel, shows correlations involving the MEN dataset. The *word2vec* baseline (solid, light-blue line) shows a rather high correlation. Averaged prototypes are slightly worse than *word2vec* embeddings (dashed, dark-blue line), again with a slight improvement at middle layers. Non-contextualised embeddings are again somewhat reliable at superficial layers, but their quality in predicting relatedness judgments deteriorates at deeper layers (with mid layers showing a somewhat less poor correlation; long-dashed bright blue line). Again we see that BERT prototypes, averaged and non-contextual, correlate with each other less and less over layers (dot-dashed, dark-red line), while averaged prototypes (two-dashed, light-red line) correlate more with the*word2vec* model than non-contextualised prototypes (dotted, red line).

Discussion

Zooming in on BERT's static embeddings, thus looking at correlations at layer -1, we see generally moderate to good correlations with similarity and relatedness ratings. Explicit rating data in general and the data sets used here specifically are often employed as the first major benchmark that computational models of semantic memory need to account for (Kumar, 2021; Mandera et al., 2017; Pereira et al., 2016). Alongside previous similar findings

(Bhatia & Richie, 2022; Lenci et al., 2022), our results thus serve as a first piece of evidence indicating that pre-trained BERT's static embeddings can offer a good implementation of semantic memory. We do however see that averaged prototypes derived from a sample of contextualised embeddings show higher correlations than the static embeddings on both tasks while non-contextualised embeddings show worse correlations at deeper layers. We thus observe that different layers in BERT's architecture capitalise on and highlight different pieces of information extracted from co-occurrence patterns and word distributions. Deriving averaged prototypes at mid layers seems to capture explicit semantic associations better, as evaluated using human ratings, in line with previous evidence that lexical semantics is best captured at central layers (Buijtelaar & Pezzelle, 2023; Rogers et al., 2020).

This first study thus shows that BERT's averaged prototype representations account well for explicit semantic ratings, on both semantic similarity and semantic relatedness. These representations are learned from running text, much like what happens with *word2vec*, one of the state-of-the-art standard methods to extract distributional semantic representations in psycholinguistic studies (Kumar, 2021; Mandera et al., 2017). However, BERT shows an advantage over a psychologically validated implementation of *word2vec* (Mandera et al., 2017) in that it better captures semantic similarity ratings. We also observe substantial correlations between relations in the BERT prototype space and *word2vec*, suggesting that the two models indeed capitalise on similar patterns. This correlation, however, decreases over layers, suggesting that prototypes derived by averaging contextual embeddings encode information that *word2vec* does not encode.

We also showed that non-contextualised representations embed informative relations only at superficial layers. Deeper layers, when faced with the lack of sentence context, quickly end up yielding a worse embedding space, which no longer captures semantic associations explicitly appreciated by human speakers. Despite their popularity (Bommasani et al., 2020), these representations are thus best avoided when modeling psychological patterns. The more interesting point, however, concerns the difference between fully static BERT embeddings and averaged prototypes. BERT learns the initial static embeddings by being exposed to several sentences and the static embedding of a word type should end up capturing a blend which, when adequately contextualised at deeper layers, affords a reliable prediction of masked words (Rodd, 2020). It is however unclear why the blends found at this layer perform worse than blends derived post-hoc by averaging a handful of randomly-selected sentences which feature the target word. The robust finding that BERT better captures semantic relations at intermediate layers might suggest that the static representation ends up capturing other aspects, including syntactic constraints, that better serve in predicting masked words (one of the two tasks on which BERT is trained) but end

up obfuscating the meaning-related aspects, which better emerge at deeper layers. Future studies should investigate in further detail how representations at intermediate layers differ from the initial, purely static embeddings to illuminate this pattern.

After showing results for explicit semantic ratings, we turn our attention to implicit semantic effects, which we investigate using lexical decision data (see also Günther et al., 2023; Mandera et al., 2017). This is an important point of difference from typical more engineering-oriented large-scale evaluations of representation systems, which exclusively focus on explicit ratings and responses (e.g., Baroni et al., 2014b; Lenci et al., 2022; Pereira et al., 2016; Vulić et al., 2020). In contrast to this, psychological theory-testing lays a strong emphasis on also investigating implicit semantic effects in processing times (Balota et al., 2007; Griffiths et al., 2007; Günther et al., 2016b; Hutchison et al., 2008; Hutchison, Balota, Neely, Cortese, Cohen-Shikora, Tse, et al., 2013; M. N. Jones et al., 2006; M. N. Jones & Mewhort, 2007; Lund et al., 1995; Marelli & Baroni, 2015; Yap et al., 2011). While one might be tempted to conclude that if BERT captures explicit semantic ratings it will inevitably also capture implicit semantic effects, there are crucial differences that warrant a dedicated analysis. First, semantic similarity ratings are relational, in that the relation between two words, both present during the decision, is evaluated by participants responding. On the other hand, lexical decision involves strings presented in isolation. Even in priming studies, where two words are presented one after another, participants respond towards the second item and are not required to consider the relation between the stimuli. Moreover, semantic similarity ratings are not a speeded response, whereas in lexical decision participants are asked to decide as quickly and accurately as possible whether a string is or is not a word. Investigating the effects of this time pressure, with reduced resources for strategic considerations, reveals automatic processing mechanisms in lexical access and semantic processing and is hence crucial for testing psychological theories of word representation. Finally, implicit measures can tap into different mechanisms than explicit ones. As we aim in this work to assess the potential of BERT's learning mechanism to yield representations that can advance our understanding of semantic memory, a full-fledged evaluation cannot stop at explicit ratings but has to investigate other types of behaviors where implicit semantic effects have been documented (Günther et al., 2023; Mandera et al., 2017; Marelli & Baroni, 2015).

Experiment 2: Primed lexical decision

We now analyse the degree to which measures extracted from BERT prototype representations are able to predict reaction times (RTs) in primed lexical decision (LD) tasks. These analyses are particularly relevant because human participants have been repeatedly shown to be influenced by lexical properties of isolated words in several psycholinguistic

tasks: while it is not necessary to postulate that participants had learned prototype representations to account for these effects of isolated words (Ambridge, 2020; Jamieson et al., 2018), we contend that a model should account for these effects if it is to be of any use as a psychologically viable account of semantics. The analysis of RTs from primed LD has also been often used to validate semantic memory models in general (Kumar, 2021) and embedding models in particular (Günther et al., 2016a; M. N. Jones et al., 2006; Mandera et al., 2017) since it offers a straightforward probe: participants are consistently faster in identifying a word as such when they are primed with a semantically more similar or related word (M. N. Jones et al., 2006; Koriat, 1981; Lucas, 2000; Neely, 1991). This analysis complements the previous one by testing whether semantic relations derived from BERT prototype embeddings capture implicit relational semantic effects, next to the explicit ratings modeled in the previous experiment.

Data

We used the Semantic Priming Project (SPP) dataset (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, Tse, et al., 2013). Participants were asked to judge whether a string is a word or not. Before seeing the target, which can be a word or a pseudoword, they are shown another word, the prime, which varies in the degree to which it is semantically associated to target words. The SPP contains RTs for primes shown at a short Stimulus Onset Asynchrony (SOA) of 200ms and at a long SOA of 1,200ms: the SOA indicates for how long the prime remained on screen. We consider both, since the influence of prime-target semantic relatedness can change depending on the time participants had to fully process the prime. We analyse RTs for 5,728 unique prime-target combinations, consisting of 1,546 unique targets and 3,048 unique primes. We discarded prime-target combinations including words for which we could not retrieve a sufficient number of sentences from the CoCA to derive the corresponding average prototype.

Methods

We derived the similarity between the prime and target word in each trial computed as the cosine between the corresponding word embeddings in the relevant embedding spaces. Higher values indicate trials where the prime and target are more semantically similar or related: we expect this predictor to have a negative effect on RTs, with shorter RTs when prime and target are more related.(Günther et al., 2016a, 2016b; M. N. Jones et al., 2006; Mandera et al., 2017).

We then first averaged RTs across subjects for *prime : target* pairs at the same SOA (in line with Gatti et al., 2022), and divided the dataset according to the SOA, short (200

ms) or long (1200 ms). We computed the Pearson's correlation between *prime : target* cosine similarities computed for *word2vec* and BERT embeddings and RTs to get a first characterization of the relation between the two. We then fitted two baseline Generalized Additive Models (one for short SOA, one for long SOA) predicting log RTs as a function of a wide range of baseline measures: target frequency, prime frequency, target OLD20, prime OLD20, target length in characters, prime length in characters, edit distance between prime and target, and Mean Binomial Frequency (MBF) of the target. Frequency estimates were retrieved from SUBTLEX-US (van Heuven et al., 2014), OLD20 was computed using the 15k most frequent words from SUBTLEX-US as reference vocabulary, while MBF was computed following the procedure outlined in Gatti et al. (2022). All predictors were then Box-Cox transformed and z-standardized separately for short and long SOA, and included in the model as simple smooths. The base model had an AIC of -12,825.05 when fitted on log RTs at the short SOA, and -13,313.68 when fitted on log RTs at the long SOA.

In a next step, we computed the cosine similarity between prime and target in the reference *word2vec* space (Mandera et al., 2017), Box-Cox transformed and z-standardized it, and added it to the baseline statistical model. In the same vein, we also computed the cosine similarity between prime and target using BERT prototypes, averaged and non-contextualised, deriving a cosine similarity measure for each *prime : target* pair at each layer, for each prototyping method (averaging and non-contextualised). We Box-Cox transformed and z-standardized also these cosine similarities and added them to the baseline GAM. For each model, we computed the ΔAIC with respect to the baseline statistical model (separately per SOA).

Results

When predicting RTs at short SOA, the model including prime-target cosine similarities from the reference *word2vec* space had an AIC of -13,353.28 with a $\Delta AIC = 330.84$ against the corresponding baseline, indicating that the additional predictor reliably improved model fit. The *prime : target* similarity had a significant effect ($edf = 4.540$, $Ref.df = 5.684$, $F = 61.433$, $p < .001$). Similarly, when predicting RTs at long SOA, the model including cosine similarities from *word2vec* had an AIC of -13,655.32 ($\Delta AIC = -195.88$) and the predictor had a significant effect ($edf = 2.321$, $Ref.df = 2.986$, $F = 68.962$, $p < .001$). The effect of *prime : target* cosine similarity in *word2vec* on RTs is negative at both SOAs: when prime and target are closer in the *word2vec* space, people are faster in recognizing the target as a word (in line with Mandera et al., 2017).

Figure 4, panel *a*, shows the correlation between prime-target similarity and log RTs separately for the two prototyping methods (color) and SOAs (faceting). The correlation

drops at first and then stabilizes between -0.1 and -0.15, but without ever approaching the correlation derived from the reference *word2vec* embedding space. Non-contextualised embeddings seem to have a stronger negative correlation with RTs in the SPP than averaged prototypes, at both SOAs. The ΔAIC (panel b) generally shows that *prime : target* similarities derived from both averaged and non-contextualised prototypes improve model fit over the baseline[9]. Averaged prototypes consistently outperform non-contextualised prototypes, and show the starkest improvement at mid-layers, in line with what we observed for explicit ratings of semantic relatedness in Experiment 1: Semantic relatedness and similarity norms. Notably, the strongest improvement in model fit for averaged prototypes is observed at the same layers for which the correlation is weaker, suggesting that at superficial and deep layers, prime-target similarities are collinear with other predictors. Pairwise correlations[10] across predictor variables computed at each layer show that the *prime : target* cosine similarity derived from non-contextualised prototypes is positively correlated with the frequency of the prime. Other predictors are only weakly correlated with the cosine similarity. Indeed, *prime : target* cosine similarities derived from averaged prototypes are weakly correlated with all other predictor variables, except for the two deepest layers, where we also see the ΔAIC decrease, confirming that the dissociation between correlation and improvement in fit depends on collinearity. However, the improvement in model fit brought by prime-target similarities extracted from the*word2vec* space is consistently stronger.

Discussion

In this analysis of implicit semantic effects on lexical decision latencies, we see that BERT prototypes perform worse than the reference *word2vec* space, and that there is a stable advantage for averaged prototypes over non-contextualised prototypes once controlling for relevant covariates. This is in line with a range of previous studies in the semantic priming literature that has played an important role in establishing the cognitive plausibility of earlier distributional word embedding models as models of semantic memory (Günther et al., 2016a, 2016b; M. N. Jones et al., 2006; Mandera et al., 2017). Our present results thus attest to the informativeness of BERT prototypes to capture psychologically relevant prime-target relations. This pattern is moreover consistent with evidence from Experiment 1: Semantic relatedness and similarity norms on ratings of semantic similarity and relatedness, although

[9] Visual inspection of GAM smooths confirm that the effects are qualitatively similar across the two and in the expected direction, with higher similarity correlating with faster RTs. The plots can be seen by running the code detailing the SPP analysis.

[10] The correlation plots can be generated using the code provided in the Online Supplementary Materials.

non-contextualised prototypes fared better with implicit semantic effects, consistently showing an improvement in model fit over a strong baseline statistical model.

Now that we have converging evidence that measure of pairwise similarity extracted from BERT capture both explicit and implicit semantic effects, although not as well as a *word2vec* space validated on psycholinguistic tasks, we turn to the analysis of semantic neighborhoods, which have been shown to influence behavior in psycholinguistics task. This effect is particularly interesting because next to being implicit is also non-relational, as it does not involve the comparison between two strings being presented to participants (simultaneously as in similarity ratings or sequentially as in semantic priming), but rather taps onto the broader organization of semantic memory.

Experiment 3: Lexical decision

After having shown that BERT prototypes can account for prime-target similarity effects in primed lexical decision, we turn to the analysis of a simple lexical decision task, where participants see a string on screen and have to quickly identify whether it is a word or not. With this data, we expect to replicate findings from previous research that words located in denser semantic neighborhoods are easier to process (e.g., L. Buchanan et al., 2001; Shaoul and Westbury, 2010; Yap et al., 2011; see also Bonandrini et al., 2023; Hendrix and Sun, 2020; Marelli and Baroni, 2015). Notably, the semantic neighborhood densities in almost all previous studies have been determined using (earlier) word embedding models, firmly placing our investigation in an established tradition of investigating semantic effects on single word processing.

We thus study here whether relations among prototype representations influence an online task, lexical decision, that has been widely studied in psycholinguistics as a window on semantic memory. Semantic richness effects in lexical decision are well documented (Pexman et al., 2008; Yap et al., 2011) and have contributed to shape theories of learning, representation and processing: if BERT fails to account for these patterns, its word representations cannot be taken as a useful approach to study semantic memory in humans without profound modifications. If semantic neighborhood density extracted from BERT, on the contrary, explains patterns in RTs, we provide a new piece of evidence that such a model may offer a fruitful implementation of semantic memory and a tool to study mechanisms involved in isolated word processing.

Data

In order to analyse to what extent BERT prototype embeddings capture variation in LD, we used the English Lexicon Project (ELP) dataset (Balota et al., 2007). Among other data,

the ELP provides RTs from a LD task, where people are again asked to judge whether a string is a word or not. Unlike in the primed LD paradigm, here participants are not shown a prime stimulus before the target decision. Using the ELP data, Yap et al. (2011) reported a facilitatory effect of semantic richness on RTs such that participants are faster in recognizing a target word when this has more semantic associates. We consider 223,380 RTs for 7,165 words from 814 participants: we only considered words which we could find in a sufficient number of sentences in the CoCA to derive the corresponding average prototype.

Methods

We computed the semantic neighborhood density (SND) of each target word in the target embedding spaces as the average cosine similarity of the 20 nearest neighbors of the target word, separately at each layer. When retrieving neighbors, we always used the same prototype method, hence the neighbors for an averaged prototype were retrieved among the averaged prototypes of the possible neighbors. The set of candidate neighbors was the same for the two prototyping methods. The reference vocabulary from which the neighbors were retrieved consisted of all the word types from the different datasets used throughout the study (SPP, ELP, MEN, and SimLex datasets, and datasets mentioned in the following studies) filtered so that at least 50 sentences could be retrieved for each target word from the CoCA in order to derive reliable averaged prototypes, for a total of 24,861 possible neighbors. Higher SND values indicate words which have denser neighborhoods in the embedding space: we thus expect this predictor to have a negative effect on RTs (Yap et al., 2011), with words from denser neighborhoods to be recognized faster.

We started by fitting a baseline statistical model using a Linear Mixed-Effect Regression (LMER) model[11] predicting log RTs from correct word trials. We included the following independent variables in the baseline statistical model.: frequency (from SUBTLEX-US, Brysbaert et al., 2012); word length in letters; OLD20 (Yarkoni et al., 2008), capturing orthographic neighborhood density as the average Levenshtein distance of the 20 nearest orthographic neighbors of the target string; Coltheart's N (Coltheart et al., 1977), capturing phonological neighborhood density as the number of words at edit distance 1 from the target in phonological space; Age of Acquisition (Kuperman et al., 2012); concreteness (Brysbaert et al., 2014); valence (Warriner et al., 2013), and word prevalence (Brysbaert et al., 2019a), all operationalized using subjective norms. All variables were accessed via

[11] Using a linear model allows us to summarise the direction of the effect as well as the improvement in model fit, to check whether next to accounting for extra variance in RTs, SND influences RTs in the expected direction or not: exploratory analyses revealed the possibility that the coefficient changes sign at different layers and a GAMM would not allow to visualize this clearly.

LexOPS (Taylor et al., 2020), were Box-Cox transformed, and z-standardized. We added random intercepts for item and participant. This model had an AIC of -356,039.3.

We then computed SND using the reference *word2vec* space (Mandera et al., 2017), and added this predictor to the baseline statistical model, after Box-Cox transforming and z-standardizing it. In addition, we derived SND for each target word at each BERT layer, Box-Cox and z-transformed it, and also added it to the baseline model (excluding the *word2vec* predictor). At each layer, we computed the ΔAIC against the baseline statistical model, and extracted the t-statistic for SND, which also gives us the sign of the coefficient. We repeated this for both prototype embeddings, averaged and non-contextualised.

Results

When adding SND derived from *word2vec* embeddings to the baseline ELP statistical model, the AIC increased by 14.656 points, indicating the extra predictor does not improve model fit. The predictor was not significant ($\beta = 0.0004$, $se = 0.0004$, $df = 6,948$, $t = 1.055$, $p = 0.291$). Considering that, in line with expectations, the Pearson correlation between SND and RTs is robustly negative ($r = -0.1008$ $[-0.1048; -0.0968]$, $t = -48.57$, $df = 229,815$, $p < .001$), this pattern is likely due to suppression effects, driven by multicollinearity: SND derived from *word2vec* is positively correlated with word frequency and negatively with AoA, both moderate correlations. Correlations with other predictors are weak.

Figure 5 summarizes the results of the statistical analyses involving BERT prototypes, with statistics derived from the reference *word2vec* space provided for comparison as a solid black line in all sub-plots. We see that the t-statistic for SND derived from BERT embeddings (panel a) starts out positive and gradually becomes negative, suggesting that SND extracted from BERT suffers less from collinearity with other covariates than SND derived from *word2vec* embeddings, whose t-statistic is very close to 0. Finally, when considering model fit expressed as the ΔAIC between the baseline model and the model icnluding SND (panel b), we again see a similar pattern for both prototyping methods and a stronger improvement in model fit at both more superficial and deeper layers. However, we have to consider that at superficial layers the correlation coefficient and t-statistic were both positive suggesting an effect in the opposite direction than we predicted, with words from denser neighborhood eliciting longer RTs. We delve deeper in to this pattern in the discussion.

Discussion

In this analysis we looked at whether BERT prototype representations capture effects of semantic richness (operationalized via semantic neighborhood density) in simple lexical

decision, to complement evidence from Experiment 1: Semantic relatedness and similarity norms about explicit semantic ratings with an analysis of implicit semantic effects and Experiment 2: Primed lexical decision about implicit effects of lexical similarity.

Patterns observed when using BERT prototypes to capture SND effects in lexical decision are more reliable then those observed when using the *word2vec* reference space - which has been validated on psycholinguistic tasks (Mandera et al., 2017). SND extracted from *word2vec*, while negatively correlated with RTs, ends up having a non-significant coefficient, due to collinearity issues. SND computed from BERT's deeper layers, on the contrary, maintains a negative effect on RTs even after controlling for other lexical predictors, in line with previous studies (L. Buchanan et al., 2001; Shaoul & Westbury, 2010; Yap et al., 2011). Moreover, especially when computed on average prototypes, SND is largely orthogonal to other variables, leading to an improved model fit from layer 9 onwards. Our results suggest that BERT embeddings might also offer a methodological improvement to derive predictors at the level of lexical semantics to be added as covariates or to be used to control for semantic relations while creating experimental stimuli, due to the absence of collinearity.

The relation between semantic richness and lexical decision times has however been at the centre of a complex debate that merits further discussion, which makes this analysis particularly interesting in the context of this work. A relevant factor concerns the degree of polysemy of a word: Rodd et al. (2002) provided evidence that homonyms, which have multiple unrelated meanings, elicit slower RTs than unambiguous words, while polysemous words, with multiple related meanings, are recognised faster. The processing cost for homonyms has been linked to the need of moving from an initial, uninformative blend state collapsing unrelated meanings, to a single state representation, matching one of the possible meanings. The advantage for polysemous words, on the contrary, has been ascribed to the emergence of strong attractors merging related yet slightly different representations. The positive correlation between SND and RTs for BERT's superficial layers which becomes negative at deeper layers should be further investigated considering this aspect (Cevoli, Watkins, Gao, et al., 2022). It is possible that at superficial layers representations tend to collapse multiple meanings, with neighborhood density being higher for words which occupy portions of the representational space close to many unrelated meanings. When contextualisation kicks in, these representations may grow sharper and more delineated, favoring neighborhood structures with multiple overlapping senses, for which a facilitatory effect of SND on RTs have been reported. Statistical models that allow to study the temporal dyanamics of predictors (Hendrix & Sun, 2020) may further illuminate this effect. We conclude by highlighting that studying how representations shift across BERT's layers is

an important aspect of the model that can illuminate non-trivial aspects of semantic memory and processing.

Isolated Word Effects: Summary

In summary, we showed that well-established measures derived from BERT embeddings (cosine similarities and semantic neighborhood densities) successfully capture a range of prominent psycholinguistic phenomena in isolated word processing (explicit ratings, semantic priming effects, semantic richness effects in simple lexical decision). Taken together, these first three studies on popular datasets thus establish BERT embeddings as a valuable source to derive prototype representations. In particular, these studies suggest that averaged contextualised embeddings should be preferred to non-contextualised embeddings. At the same time however, they also show that static embeddings at the superficial layer already encode informative lexical representations that closely mirror those learned by static DSMs like *word2vec*. However, BERT prototypes do not encode as informative semantic relations as a state-of-the-art *word2vec* model validated on psycholinguistic tasks (Mandera et al., 2017) when it comes to ratings of semantic relatedness and prime-target relations in primed lexical decision.

Up to this point, we have treated BERT as if it were another static word embedding model in a long line of such models (for overviews, see Günther et al., 2019; Kumar, 2021). While we have defined the validity of type-level semantic representations as a necessary condition for a model of semantic memory, we have not yet tapped into the crucial component that sets models like BERT apart from their static predecessors, and which is central to our present endeavor: The possibility to derive context-sensitive meaning representations for each individual word token in a sentence. Thus, having established an empirical foundation at the type-level, we can now more confidently venture in further analyses involving the relation between prototypes and contextualised token-level representations.

Replacements

After establishing the robustness and informativity of prototype (i.e., type-level) representation) extracted from BERT, we turn to examine to what extent context-free lexical semantics on the one hand and context information on the other hand determine meaning representations of word *tokens* in natural sentence contexts. To this end, we extracted sentences containing specific target words from the CoCA, and systematically varied the target words (by replacing them with more or less related words) while keeping the rest of the sentence context constant. For example, when selecting the sentence *The lion hunts*

down the gazelle. for the target *lion*, this target could be replaced with an unrelated word, e.g., *bathtub*, resulting in *The bathtub hunts down the gazelle.* or with a related word, e.g., *tiger*. We sample replacements at different levels of relatedness, operationalised using similarity derived from free semantic associates from the English Small World of Words (SWoW) dataset (De Deyne et al., 2019). This experiment is similar in spirit to the one conducted by Ettinger (2020) on event knowledge and semantic role sensitivity in BERT, but considers more graded relations and extends the investigation to similarity across targets and replacements as well as to their semantic neighborhoods, rather than focusing on which words BERT predicts in given contexts, to probe the underlying representational space rather than the output of the model predictions.

This setup allows us to test if BERT representations meet requirement 2 (appropriate context-sensitivity of token representations): if the context has an effect, the token representations for targets and replacements should, over BERT's layers, diverge from their respective type representations and, since they appear in the same context, converge towards one another. At the same time, it also allows us to test whether BERT representations satisfy requirement 1a (non-arbitrary word use): there should still be a difference between the token representations for the different replacements, to the extent that the original target fits the context better than the replacements, and that the token representations are more similar to their respective type representations than to the replacements' type representations. Furthermore, by systematically varying the degree of semantic relatedness between targets and replacements (for example, replacing *lion* with *leopard*, *cat*, *bear*, and *bathtub*), we can formulate these expectations in a more precise and graded manner, as will be outlined in the different simulations below.

Experiment 4: Lexical representations in context

We begin our exploration of the interaction between contextual influence and lexical semantics by analysing the lexical representations of words in context. To this end, we compare the representation of a target word in a sentence in which it actually has been used in a representative corpus against that of a replacement word, varying the degree to which replacements are related to the original word (in a psychological model of lexical semantics derived from free semantic associations; De Deyne et al., 2019).

Data

We first identified a set of target words, starting from the word types available in the Small World of Words (SWoW) dataset (De Deyne et al., 2019). We filtered this set only considering relatively frequent words with a Zipf score (Brysbaert et al., 2019a) ≥ 3 in

SUBTLEX-UK (Van Heuven et al., 2014), and with a percent of speakers knowing the word $\geq 50\%$ in the prevalence norms by Brysbaert et al. (2019b). We restricted our focus on nouns, verbs, and adjectives as indicated in SUBTLEX-UK (both accessed via LexOPS, Taylor et al., 2020), obtaining 5,683 nouns, 1,514 verbs, and 1,738 adjectives for a total of 8,936 word types. We then retrieved all the sentences from the CoCA which contain one of the target words. After discarding 10 word types for which we could not retrieve at least 50 sentences, we were left with a set of 8,926 word types.

Experiment 4a: Pairwise similarities of targets and replacements

Methods

For each of the 8,926 target word types, we derived the cosine similarity with all the other word types using the pairwise similarity relations computed using the random walk algorithm applied to the free association data from the English dataset in the SWoW project (SWoW-EN; De Deyne et al., 2019). We considered the three associates provided by each participant for a given cue word. This algorithm implements a spreading activation mechanism and creates a global measure of similarity from behavioral responses to a key psychological task such as producing free associates given a cue word. We take the resulting similarities as our gold-standard: terms that are similar in this space are taken to reflect the fact that native English speakers see them as related, consistently with empirical evidence that similarity metrics derived from this space account for several benchmarks, outperforming other mechanisms to extract similarity from free associates. We used code from De Deyne et al. (2019), keeping the hyper-parameter values they recommend and refer the interested reader to the original publication for the technical details of the algorithm used to derive pairwise similarities. This step caused the loss of an extra word, leaving us with a set of 8,925 word types.

For each word type, we then ranked all other word types by cosine similarity. For each target word type we then sampled 5 replacement words, considering the following bins: the most related word type (rank $\in]0; 1]$), closely related words (rank $\in [2; 10]$), related words (rank $\in [11; 50]$), loosely related words (rank $\in [51; 100]$), unrelated words (rank > 100). Replacements were sampled across word types with the same PoS tags to ensure that replacements do not violate loose syntactic restrictions when substituting a target word in the sentence it occurs in. As an example, consider the target word *cathedral*, for which we sampled the candidate replacements *church* (cosine similarity = 0.525), *abbey* (cosine similarity = 0.384), *congregation* (cosine similarity = 0.367), *chateau* (cosine similarity = 0.171), *brewery* (cosine similarity = 0.104) (listed from the closest to the farthest).

After having sampled 5 replacements for each target, we retrieved 50 sentences from the CoCA (limiting the sentence length to the maximum context size that BERT can handle, i.e., 512 tokens), and replaced the first occurrence of the target word in each of them (see Table 1 for an example of how sentences were constructed). We then fed each sentence to our BERT embedder to obtain contextualised vector representations for the target word, as well as for the 5 replacements, at every layer. We investigate the similarity between lexical representations of targets' and replacements' contextualised and prototype representations, obtained following the pipeline described in the section Token and Type representations.

Word	Replacement	Sentence
cathedral		Yesterday, we went to the **cathedral** for a christening.
	church	Yesterday, we went to the **church** for a christening.
	abbey	Yesterday, we went to the **abbey** for a christening..
	congregation	Yesterday, we went to the **congregation** for a christening.
	chateau	Yesterday, we went to the **chateau** for a christening.
	brewery	Yesterday, we went to the **brewery** for a christening.

Table 1

Example of the generation of inputs for the replacement studies. We started from a sentence from the CoCA containing the target word - ***cathedral*** *in this example. We then sampled 5 replacements with the same PoS tag at different levels of similarity in the English Small World of Words dataset (SWoW-EN) similarity matrix. Finally, we replaced the target word in the original sentence with each of the replacements, yielding 5 sentences with progressively more severe semantic incongruences. The sentence displayed here serves as an example: actual sentences are typically much longer.*

For each sentence where we replaced a word, we thus have the original target and the different replacements, for which we can access contextualised token-level representations as well as prototype (type-level) representations, whose similarity relations we want to investigate. Specifically, we obtained the following representations:

- contextualised representations of the target ($target_c$), obtained from the original sentences from the CoCA. These representations encode the token-level semantics of the target word in contexts in which it is observed in a representative corpus of English.
- contextualised representations of the replacement ($replacement_c$), obtained from the sentences extracted from the CoCA where the target word was substituted with each of the replacements. These representations encode the token-level semantics of each replacement word in a context where it has not been observed in our sample.
- average prototype representations of the target or the replacement ($target_{p-avg}$, $replacement_{p-avg}$), obtained by averaging the token-level representations of the target

or the replacement when occurring in 50 original sentences from the CoCA. These representations encode the type-level lexical semantics of each word – target or replacement –, considering a sample of the contexts in which it occurs in a representative sample of English (see Experiments 1–3).

- non-contextualised prototype representation of the target or the replacement ($target_{p-nc}$, $replacement_{p-nc}$), obtained by feeding the target or the replacement word to the BERT embedder in isolation. These representations encode the type-level lexical semantics of each target word, operationalized as the representation of a word outside of any sentence context (see Experiments 1–3).

Given these representations, we derived the following cosine similarities from BERT representations, which we compare to "psychological ground truth" SWoW-EN similarity relations:

1. $cos(target_{p-[avg|nc]}, replacement_{p-[avg|nc]})$ (the shortcut $p-[avg|nc]$ denotes both $p-avg$ and $p-nc$, so this term describes two different measures): the cosine similarity between the prototype representations of target and replacement. In line with correlations with MEN and SimLex reported in Experiment 1: Semantic relatedness and similarity norms, we expect targets and replacements found to be closer in the SWoW similarity space to be also embedded closer by BERT, at various layers.

2. $cos(target_c, replacement_c)$: the cosine similarity between the contextualised representations of target and replacement. We expect that replacements that are farther from the target in the SWoW similarity space will be harder to integrate in the sentence context, resulting in less similar token-level representations of the replacement and the target.

3. $cos(target_{p-[avg|nc]}, replacement_c)$: the cosine similarity between the contextualised representation of the replacement word and the prototype of the target word. Again, we expect these similarities to mirror those in the SWoW similarity space, as the prototype target vector encodes the target lexical semantics while the replacement word is embedded in a sentence where it does not fit. If context sufficed to shift the replacement's token-level representation close to the target prototype regardless of how scarce the semantic similarity between the replacement and the target is, it would be evidence that the semantic core learned by BERT would be too easily overridden and that any word in context would be interpreted simply based on the context.

4. $cos(replacement_c, replacement_{p-[avg|nc]})$: the cosine similarity between the contextualised and prototype representations of the replacement words. The contextualised representation is extracted from a sentence where the token is not originally found (since it replaces another word). The prototype representation, on the contrary, supposedly captures the lexical semantics of the word regardless of context. We can therefore examine how the sentence context affects the token-level representation, and we expect that the two representations are similar at superficial layers but diverge at deeper layers, where context will have shifted the contextualised representation of the replacement towards the semantics of the original target. This shift is expected to be stronger for replacements that are farther away from the target, since we can expect context to exert a stronger influence in these cases.

We are interested in the extent to which relative relations of proximity observed in SWoW similarity patterns predict pairwise similarities computed from BERT embeddings at different layers, and therefore fit a series of regression models where BERT similarities are the dependent variable and SWoW similarities are used as the independent variable in interaction with BERT's layer.

We first averaged the BERT cosine similarity measures for each target-replacement pair over the 50 sentences in which a replacement substituted the target. Then, we applied a Box-Cox transformation and z-standardization to cosine similarity measures separately per layer, so that each cosine similarity measure at each layer was centered around 0: negative values would indicate lower similarity. Transformation and standardization were applied to counter the known anisomorphism of BERT embeddings, which becomes more and more pronounced at deeper layers (Ethayarajh, 2019).[12]

We fitted different GAMs, each predicting a different cosine similarity in BERT as a function of the cosine similarity in the SWoW similarity space (included as a simple smooth), the layer (also included as a simple smooth), and a partial tensor product between $SWoW$ cosine similarity and layer. The first model predicted $cos(target_p, replacement_p)$; the second targeted $cos(target_c, replacement_c)$; the third considered $cos(target_p, replacement_c)$, and the fourth addressed $cos(replacement_p, replacement_c)$. We fitted separate models for prototype vectors derived by averaging contextualised representations and by presenting non-contextualised words. We expect similarities from the SWoW space to reliably predict

[12] Essentially, the cosine similarity across BERT embeddings becomes larger between any two pair of words at deeper layers, since representations occupy a narrower and narrower hyper-cone of the space. The same pair of unrelated words, e.g. *dog : fog*, would have a numerically small similarity at layer 0 which would however grow at deeper layer, without this necessarily meaning that the two words grew closer relative to other word pairs.

cosine similarities in BERT at all layers. We chose to fit GAMs to explore the possibility that the relation takes a non-linear form, rather than assuming a linear relation. Given that we z-standardized within each layer, the layer variable by itself should have no effect; however, the partial tensor product can reveal whether the relation between BERT and SWoW similarities changes over layers, and thus depends on the (non-)success of the contextualization process.

Results

We first see that BERT prototypes preserve relations of similarity found in the SWoW similarity space (Figure 6, although differences emerge between averaged and non-contextualised prototypes and we see a general worsening of the fit at deeper layers. Much like averaged prototypes fitted relatedness and similarity ratings (Experiment 1), they also preserve relations derived from free associations, again with a slight advantage at mid layers. However, we see that at layer 12, the deepest one, similar words in SWoW are not embedded in a way that preserves relative similarity in BERT. We get back to this pattern in the discussion. Non-contextualised prototypes show a good fit at superficial layers, which however degrades very quickly until no clear relation is observed at layer 12.

Figure 7 shows the relation between similarity in the SWoW similarity space and the BERT similarity between target and replacement contextualised embeddings; we present a single plot since no prototyping is involved. Here we see that there is a fundamental coherence between BERT and a psychologically grounded model of meaning derived from free association data (De Deyne et al., 2019). The plot exhibits a remarkable fit between cosine similarities derived from the two semantic models, one behavioral and the other computational. This shows that contextualisation does not override differences in lexical semantics entirely: If the token-level BERT representations were fully determined by context, then the contextualized target and replacement representations would have become virtually identical and would thus *not* mirror the SWoW semantic similarity structure. Instead, the relation between the two models is remarkably consistent at all BERT layers, from the early and still relatively context-independent layers to the late and heavily contextualized layers.

Figure 8 shows the relation between SWoW similarities and $cos(target_p, replacement_c)$, separately for the two prototyping methods (left: average prototypes; right: non-contextualised prototypes). We see that averaging yields rather consistent relations between BERT and SWoW over all layers, with high similarity resulting from SWoW's free associations patterns predicting high relative similarity in BERT (with the curves flattening at very high SWoW similarities). When deriving BERT prototypes by feeding isolated words we see a positive relation at superficial layers which however largely

disappears at the deepest layer, a pattern we have seen in all analyses comparing semantic representations so far. This strengthens the interpretation that when BERT has no context to integrate and is asked to perform a task which is quite far from its training objective, its representations no longer capture semantic relations robustly.

The difference between averaged and non-contextualised representations is also found when considering $cos(replacement_p, replacement_c)$ (see Figure 9). Here we expect to see an indirect effect of the similarity between target and replacement in the SWoW similarity space. Consider the sentence *The pride of lions chased the gazelle* and consider replacing the target *gazelle* with a semantically similar word, *rhino*, and a dissimilar word, *fork*. At the early layers, $replacement_p$ and $replacement_c$ are always very similar, because context has not yet affected these embeddings. As this is independent of the relation between target and replacement, we expect no relation between similarities in BERT and similarities in SWoW at the early layers. At deeper layers, however, the context (in which the replacement is normally not found) will exert more and more influence on $replacement_c$: if the replacement is very dissimilar than the target (as with *fork*), the context should shift the replacement's token-level representation further away from the replacement prototype and closer to the target prototype (which in turn leads to $cos(replacement_c, replacement_p)$ becoming more aligned with $cos(target, replacement)$ from SWoW). If the replacement already is very similar to the target, this should happen less.

In Figure 9, we see that the expected relation between BERT and SWoW similarities emerges over layers at least up to a certain similarity score in the SWoW similarity space, whereas the effect appears flat around 0 at the lower end of the similarity values in the SWoW semantic space. This indicates that contextualisation starts exerting a relevant influence already after only a few layers have been traversed, implying that context starts playing a role soon in BERT's architecture. Therefore, the contextualized representation of the replacement changes through BERT's layers and becomes more and more similar to that of the target, as reflected in the fact that the SWoW similarity between target and replacement predicts the similarity between the contextualised and prototype representations of the replacement. This pattern attests to a systematic and meaningful influence of context on lexical representations in BERT. Moreover, we see that this pattern emerges only when the target and the replacement are rather similar in SWoW. When this is not the case, BERT struggles to morph the replacement's contextualised representation into one resembling the target. This again confirms the role of a word's semantic core, which influences the degree to which it can or cannot be contextualised successfully.

In the next experiment we go beyond comparing individual representations and additionally consider their semantic neighborhoods. This will allow us to better understand

where, in the underlying semantic space, representations shift through contextualisation, and to better characterise the influence that contextualisation exerts on the static representations. We will discuss the results of these two related experiments at the end of the section.

Experiment 4b: Neighborhood structure

The previous analysis remains mostly blind as to how exactly contextualisation unfolds and how meaning representations shift with respect to the larger semantic space when context exerts its influence. Even if they preserve relative pairwise similarities, for example, the contextualised representation of the replacement could move away from its corresponding prototype because the contextualised representation becomes meaningless. For example, *gazelle* and *fork* might be dissimilar because the context *A pride of lions chased the [word].* pulls the embedding of *fork* towards totally unrelated words to both *gazelle* and *fork*. If this is the case, we contend that contextualisation takes place but still fails, since we expect the context to influence *fork* in a systematic way. However, pairwise similarity patterns described above might result from contextualisation pulling the embedding of the replacement (*fork*) closer to the representation that the target *gazelle* gets in the same sentence. By comparing the semantic neighbors of the replacement and the target in the same sentence, we can check to what extent the former representation gets dragged towards the latter by context, and get a more detailed picture of how representations shift in the embedding space as a function of their context-independent similarity and the level of contextualisation (operationalized as the BERT layer).

Data

We used the same targets, replacements, and sentences as in Experiment 4a: Pairwise similarities of targets and replacements.

Methods

Following a method introduced in the study of language change (Gonen et al., 2021) to detect words whose meaning changed over time, we rely on the Jaccard coefficient (a standard measure for semantic neighborhood comparisons; Cassani et al., 2021; Hamilton et al., 2016), which is computed as the ratio between the cardinality of the intersection between two sets and the cardinality of their union. Higher coefficients indicate a higher set overlap, and thus a higher consistency of the neighborhoods. Unlike cosine similarities, Jaccard coefficients are not affected by the anisomorphism of BERT spaces, removing the need for further transformations and making the results more interpretable at face value.

We started by retrieving the 250 nearest neighbors at each layer for the four target representations ($target_p$, $target_c$, $replacement_p$, $replacement_c$). Neighbors were retrieved among the prototype representations at the same layer and ranked by cosine similarity. Prototype representations exist in the same embedding space as the contextualised representations and can thus be directly compared. Neighbors of contextualised representations were retrieved separately from averaged and non-contextualised prototypes, whereas neighbors of prototype representations were derived among prototypes obtained in the same way, so the neighbors of an averaged prototype can only be other averaged prototypes. The set of possible neighbors was the same we used to compute SND, consisting of 24,861 word types.

We then computed the following metrics to capture the coherence of semantic neighborhoods:

- $jaccard(target_c, replacement_c)$: the Jaccard coefficient between the 250 nearest neighbors of the contextualised target and the 250 nearest neighbors of the contextualised replacement. The higher the Jaccard coefficient, the more neighbors are shared and thus the more the two neighborhoods are consistent with each other.

- $jaccard(replacement_c, target_{p-[avg|nc]})$: the Jaccard coefficient between the 250 nearest neighbors of the contextualised replacement and the 250 nearest neighbors of the prototype target.

- $jaccard(replacement_c, replacement_{p-[avg|nc]})$: the Jaccard coefficient between the 250 nearest neighbors of the contextualised replacement and the 250 nearest neighbors of the prototype replacement.

- $\Delta jaccard = jaccard(target_{c-1}, target_{c12})$ - $jaccard(replacement_{c-1}, replacement_{c12})$: the difference between (i) the Jaccard coefficient between the 250 nearest neighbors of the contextualised target at layer -1 (the fully static embedding) and layer 12 (after context has been fully integrated in the token representation), and (ii) the Jaccard coefficient between the nearest neighbors of the contextualised replacement at layers -1 and 12. We did this for every sentence in which the target appears, with the prototype representations of all targets and replacements at layers -1 and 12 respectively serving as possible neighbors. A positive number indicates that the target's neighbors were more coherent than the replacement's neighbors between layers -1 and 12.

For the first two Jaccard coefficients, $jaccard(target_c, replacement_c)$ and $jaccard(replacement_c, target_{p-[avg|nc]})$, we expect the coefficient (i) to increase at deeper layers, which would indicate that context has indeed shifted the replacement representations

towards that of the target (contextualised and prototype), and (ii) to drop when similarity in the SWoW similarity space also drops, showing that it is harder to integrate more dissimilar words in context. On the contrary, we expect $jaccard(replacement_c, replacement_{p-[avg|nc]})$ to be high at superficial layers, where context has not yet affected the contextualised representation of the replacement, which is thus close to the corresponding prototype, but lower at deeper layers, where context has shifted the representation towards that of the target. For the latter reason, we also expect the coefficient to drop when similarity in the SWoW similarity space drops, consistently with the other two coefficients.

Finally, concerning $\Delta jaccard$, we expect the target to have a higher coherence between neighbors at layer -1 and 12. On the contrary, the replacement's neighbors should differ more between superficial and deep layers, as the sentence context will drag the replacement's representation away from its prototype representation (hence, its semantic core) to look more like the target. We expect, therefore, that $\Delta jaccard$ increases when target and replacement are farther apart: While the target-component of the difference should remain unaffected, replacements that are farther from the targets they replace are expected to be more affected by context and thus move farther away from their prototype (resulting in smaller Jaccard coefficients for the replacement-component and thus a larger overall Δ).

We once again averaged Jaccard coefficients for the same *target : replacement* pair over the 50 sentences in which a same replacement substituted the target word, and fitted a GAM to predict each Jaccard coefficient as a function of layer (simple smooth), cosine similarity between target and replacement in the SWoW similarity space (simple smooth), and a partial tensor product between the two.

Results

Figure 10 shows the predicted coherence for the first three Jaccard coefficients ($jaccard(target_c, replacement_c)$: top panels; $jaccard(replacement_c, target_{p-[avg|nc]})$: mid panels; $jaccard(replacement_c, replacement_{p-[avg|nc]})$: bottom panels) as a function of similarity between target and replacement in the SWoW similarity space, separately for the two methods to derive prototypes (averaging: left panels, non-contextualised: right panels). $jaccard(target_c, replacement_c)$ increases with the increase in SWoW similarity as well as at deeper layers with a significant interaction, in line with our predictions. This shows how the contextualised embedding of the replacement gets pushed closer to the contextualised embedding of the target when the target-replacement similarity increases and when context exerts a stronger influence. This confirms that the embedding's replacement changes in the direction of the semantics presupposed by the target it replaces (Nieuwland & Van Berkum, 2006). It further demonstrates that the contextualisation process yields increasingly similar

results to the original target when the prototype embedding of the replacement is already quite similar.

$jaccard(replacement_c, target_{p-[avg|nc]})$ presents a similar pattern when considering averaged prototypes, but unexpectedly flips when looking at non-contextualised embeddings. In this case, the coherence decreases at deeper layers, suggesting that when the prototype is derived with no context, the representation moves away from the target and indicating once again that non-contextualised prototypes deteriorate at deeper layers due to the lack of sentence context. This analysis complements the previous one by showing that the contextualised replacement does not simply shift closer to the representation of the target in the same context, but also closer to the prototype representation of the target, at least when considering averaged prototypes, and further attests to the systematic (and expected) interaction between lexical semantics and context.

In the bottom panels, we see how $jaccard(replacement_c, replacement_{p-[avg|nc]})$ changes over layers. Unsurprisingly, at layer 0 the SWoW similarity does not influence neighborhood coherence, since the neighborhoods being compared pertain to the same word type, and the coherence is rather high. When moving to deeper layers, however, we see the coherence drops substantially, for both prototype approaches (more for non-contextualised prototypes). SWoW similarity plays a very minor role at all layers, with slightly more coherent neighborhoods between the contextualised and prototype replacement at higher SWoW similarity values, in line with expectations. This confirms that when a target is replaced by a word which is semantically more related, the neighbors of the contextualised replacements tend to overlap more with those of the target, although the main influence here is that of contextualisation over layers, which pulls the replacement in context away from its lexical semantics. On the basis of this analysis, along with those presented in Experiment 4a: Pairwise similarities of targets and replacements, non-contextualised embeddings at deeper layers confirm to provide unreliable representations for semantic analyses.

The last analysis we present in this section quantifies the difference in coherence between the neighbors of the contextualised target and the contextualised replacement between layer -1 (encoding static embeddings) and layer 12 (encoding fully contextualised embeddings). Figure 11 shows the estimated effect: As indicated by the blue solid line reflecting averaged prototypes, increases in similarity between target and replacement in the SWoW similarity space come with a decrease in the Jaccard Δ ($edf = 7.698$, $Ref.df = 8.57$, $F = 120.6$, $p < .001$), eventually reaching 0 at high SWoW similarity values. For averaged prototypes we thus see that when the replacement is semantically dissimilar from the target in the base psychological model of meaning (i.e., low values on the x-axis in the plot) contextualisation drags the representation further away from its semantic core than it does

for the target. When the replacement is similar to the target, and thus is expected to fit better in the sentence, the two contextualised representations are affected in a similar way and are embedded consistently between the static layer and the fully contextualised layer.

On the contrary, the difference is 0 regardless of SWoW similarity when considering non-contextualised prototypes ($edf = 1.006$, $Ref.df = 1.012$, $F = 0.02$, $p = 0.907$), denoting no relation between neighborhood coherence and prototype similarity in a psychological model of meaning. This, once again, confirms how the embedding space derived from non-contextualised words is a poor reflection of semantic relations.

Discussion

This set of analyses shows that BERT's embeddings encode lexical semantics in a way which is highly consistent with the SWoW similarity space as probed using both pairwise similarities and neighborhood coherence. The pattern already emerges at the static embedding layer (-1 in the plots) but also once contextualisation unfolds. Thus, BERT preserves a semantic core of words even through its contextualisation, meeting our first desideratum. Moreover, in line with previous evidence from Experiment 1: Semantic relatedness and similarity norms and Experiment 2: Primed lexical decision, we see that at deeper layers, non-contextualised prototypes pay a price in the fidelity of their representations (as compared to averaged prototypes). It is worth thinking of the objective BERT is trained with, predicting a masked word from its context: The embedding should support an accurate prediction of the word type that occurs in place of the mask. When we replace a target word with a very different word (left end of the x-axis in the plots shown for the last two experiments), contextualisation struggles to integrate such a semantically different replacement (due to the static embedding encoding a word's lexical semantics) and the resulting representation is very different from that of the target prototype (see, for example, Figure 9). However, when the similarity between the target and the replacement in the SWoW similarity space increases, the replacement can be more easily contextualised, as evidenced by the higher neighborhood coherence between the neighbors of the contextualised replacement and those of the target (see Figure 10, top and mid row). This pattern captures the vital interaction between lexical semantics on the one hand and context on the other hand. Critically, BERT's context-sensitive architecture allows the model to learn this interaction in a data-driven manner, enabling it to successfully shift a word's semantic representation based on context it appears in.

Experiment 4b: Neighborhood structure further confirms that while context influences token-level representations in BERT in systematic and meaningful ways, words cannot be arbitrarily replaced in context, meeting the desiderata for requirements 1b and 2

stated in the Introduction. Contextualisation – the influence exerted by context at deeper layers in the network – shifts the representation in meaningful ways, as captured by the neighborhoods: the in-context replacement will be pulled away from its respective prototype replacement and closer to the target representation. Finally, contextualisation is predictably more successful when the semantic cores of target and replacement are more similar.

This context sensitivity is important: a newspaper should look less like a prototypical newspaper in the sentence *I used a newspaper to cover my head when it started raining* (and, for example, provides very different affordances; Glenberg and Robertson, 2000). Nieuwland and Van Berkum (2006) show that in sufficiently supportive discourse contexts, animacy-violating predicates (e.g., *the peanut was in love* presented after a story about a peanut with anthropomorphic features) elicited no N400 components whereas predicates relating to the core meaning of the word form (*the peanut was salted*) did result in perceived anomalies. The authors take this as evidence that language comprehension does not involve an initially context-free semantic analysis. Anectodal evidence from Cevoli, Watkins, Gao, et al. (2022) suggests that when presenting BERT with the discourse prompts used in the original study, the model predicts words such as *singing, talking,* or *dancing* after *The peanut was*. This indicates that the model's representation of peanut at deeper layers - those that support the prediction of masked words - has shifted away from the prototypical representation of *peanut* and closer to that of *person*. However, this does not necessarily rule out the presence of an underlying stable, static, context-independent semantic representation. Our analyses generalise the exploration conducted by Cevoli, Watkins, Gao, et al. (2022) to many different targets and replacement with graded relations of semantic similarity to the target. Even though our contexts consisted of a single sentence, we still see that context interacts with the static embedding in informative ways, which manifest in how contextualised representations shift in relation to prototype representations. Our results thus speak to the co-existence of context-independent and context-dependent representations: in particular, we see that BERT can in principle account for semantic violations that are determined by context and go against semantic violations entailed by the context-independent semantics of concepts (Nieuwland & Van Berkum, 2006).

Based on our results one could also expect that the lower the semantic similarity between the context-independent representations of target and replacement, the richer the context should be to over-ride the expected semantic anomalies. This hypothesis could be further tested by replicating Nieuwland and Van Berkum (2006)'s experiment varying the informativity and length of the discourse preceding the target violation. BERT thus presupposes a model of semantic processing which combines context-independent representations, whose role emerges clearly from the wealth of semantic effects for isolated

words (some of which we addressed in the first section of this work), with context-dependent representations that emerge considering the representations themselves as well as their neighborhoods.

The analysis we performed can also speak to theories of metaphor (Kintsch, 2000; Lakoff & Johnson, 1980), for which distributional spaces have already been fruitfully used and which have long served as a prominent test case for context-sensitive distributional models (Harati et al., 2021; Kintsch, 2000; Utsumi, 2011; see Nick Reid and Katz, 2018). We show that BERT can offer a model of how new metaphors could come to be, potentially predicting which words could function better as vehicles for certain tenors by considering how well they can be integrated in context and how their nearest neighbors shift with respect to the semantic field being targeted.

After having established that BERT accounts for isolated word effects and that the representation it learns integrate context effects with a stable semantic core, we turn to investigate context effects in more detail. Specifically, we consider how meaning changes as a function of context when different senses are involved in order to analyse the nature of meaning shifts.

Meaning in Context

So far we have shown that prototype representations capture offline (similarity and relatedness ratings) and online (semantic neighborhood and prime-target similarity effects in lexical decision tasks) psycholinguistic semantic effects. Moreover, we saw that the BERT model is sensitive to the interplay between prototype and context-dependent representations by replacing certain words in context with other words at varying degrees of semantic similarity as gauged from a psychological model of meaning (De Deyne et al., 2019). In the last suite of experiments we address the very same question from the opposite perspective, keeping the target words constant while changing the sentence contexts. This will provide us further evidence with respect to requirement 2 (appropriate context-sensitivity of token representations) and requirement 1a (non-arbitrary word use). We also investigate whether BERT preserves meaning differences reflected in lexical taxonomies, by investigating how it captures different senses of homonyms and polysemous words and whether it even captures more fine-grained meaning differences that do not necessarily match different senses (cf. the examples we provide in the introduction about *chairs* and *tomatoes*, where the same word in different sentences refers to the same sense or even the same exact referent, but evokes very different associations, affordances, and semantic features).

To this end, we leverage three datasets that allow us to investigate how representations change as a function of context when considering words whose *meaning*

changes as a function of context: (i) the *Words in Context* (WiC) dataset (Pilehvar & Camacho-Collados, 2018); (ii) the stimulus materials from the seminal study by Till et al., 1988, who conducted a priming experiment with polysemous primes introduced in disambiguating sentences; and (iii) the *Meaning Modulations* (MeMo) dataset, which we developed for the purpose of this work. This dataset goes beyond the issue of polysemy, instead focusing on nuanced differences in meaning for words used in the same sense but that nonetheless undergo slight modulations due to context, such that they would elicit different associations.

In studying how context modulates meaning across and within sense boundaries, we compare contextualised embeddings from BERT with the Instance Theory of Semantics (ITS), a purely exemplar-based model proposed by Jamieson et al. (2018), which was shown to produce representations that support semantic effects pertaining to isolated words as well as capture different senses of homonyms.

Experiment 5: Word Sense Disambiguation

The *WiC* dataset (Pilehvar & Camacho-Collados, 2018) is a popular benchmark in the NLP community to test algorithms for Word Sense Disambiguation (WSD): we use it to study whether the representation of a same word type changes following sense distinctions reflected in sentence context. As an example, consider the word *bark* in sentences *The bark of the tree was slightly burnt.* and *We heard the bark of a dog in the distance.*. For BERT to offer a useful model of meaning for psychological purposes, we expect the representation of *bark* in the two sentences to grow different over layers and to become more similar to that of possible synonyms in each context, e.g., *cortex* or *howl* respectively. While previous work has extensively addressed BERT's ability to reflect context-dependent meaning shifts (see Apidianaki, 2023; Haber & Poesio, 2024, and references therein for a comprehensive summary of how these models have been used to detect different word senses in NLP), our analysis here focuses on how the process unfolds over layers and especially on the relation between context-dependent representations of appropriate and inappropriate synonyms in context. Given an ambiguous word, like *bark*, both *cortex* and *howl* could be valid synonyms of the word in isolation: given the evidence presented about replacing related words in context, we zoom in here on what happens in BERT's representations when two possible sense-dependent synonyms replace an ambiguous word in a context where the synonym is indeed a synonym of the ambiguous word, and in a context where it is not.

Data

The WiC dataset provides polysemous words, each used in two different target sentences. The original task for which the dataset was created is to determine whether a given target word is used in the same sense in the two sentences or not. As an example, consider the word *ball* and two target sentences *The player swung but missed the ball.* and *The girl dressed up to attend the ball.*. In this case, the dataset indicates that the target word is used in two different senses. On the contrary, for the target word *beat* and the sentences *We beat the competition.* and *Agassi beat Becker in the tennis championship.*, the gold standard indicates the same sense. Human performance puts the upper bound for this dataset at 80%, with BERT scoring above 70% in the task of WSD (Breit et al., 2020), indicating promising performance. Rather than testing BERT on the task again, we aim to investigate how the representations of the target word shifts across layers with respect to other words which are semantically related to each sense of the target. We thus provide a more fine-grained and semantically more sensitive analysis than studies on pure word sense disambiguation.

Methods

We first filtered only the items where the target was used in a different sense, since we are interested in how the representation of a word shifts when used to index different meanings. Then, we manually annotated a sample of items by adding a synonym of each sense from its respective BabelNet synset[13]. Going back to the *ball* example introduced before, we annotated the first sentence with the synonym *sphere* (the *appropriate* synonym for the first sentence and the *inappropriate* synonym for the second) and the second with the synonym *party* (the appropriate synonym for the second sentence and the inappropriate synonym for the first).

We had 92 different target words in our final annotated dataset, each used in a variable number of sentences between 1 (the vast majority of target words, 56) and 12 (the highly ambiguous word *head*). 143 different words were used as the synonym of the target word in the first sense. 138 different words were used as the synonym of the target word in the second sense. These statistics show a considerable variability in the number of target words and senses, offering an informative benchmark to assess the ability of computational models to tease apart different senses of ambiguous words and, crucially, relate those senses to appropriate synonyms resulting from expert annotations. Since each item is embedded at the usual 14 BERT layers, we worked with 2,548 data points in the statistical analysis.

[13] We release this resource with our annotations for replication purposes in the OSF repository.

We derived the following embeddings from BERT, extracting token-level representations at the usual 14 layers:

- $target_{s1|s2}$: the contextualised embedding of the polysemous target word (*ball* in the example) in the first ($s1$) and second ($s2$) sentence;
- $appropriate_{s1}$: the contextualised embedding of the appropriate synonym of the target in the first sentence, *sphere* in the sentence *The player swung but missed the {WORD}.*;
- $appropriate_{s2}$: the contextualised embedding of the appropriate synonym of the target in the second sentence, *party* in the sentence *The girl dressed up to attend the {WORD}.*;
- $inappropriate_{s2}$: the contextualised embedding of the inappropriate synonym of the target in the second sentence, *party* in the sentence *The player swung but missed the {WORD}.*;
- $inappropriate_{s1}$: the contextualised embedding of the inappropriate synonym of the target in the first sentence, *ball* in the sentence *The girl dressed up to attend the {WORD}.*;

From these representations, we derived the following metrics:

- $cos(target_{s_1}, target_{s_2})$: the cosine similarity between the contextualised embeddings of the target word in the two sentences. We expect these representations to diverge over layers, reflecting the fact that before context is integrated, they point to the same location in semantic memory, but once context is factored in, they identify different portions of semantic space, in line with the two different senses.
- $cos(target_{s_n}, appropriate_{s_n})$: the cosine similarity between the target's contextualised embedding in a sentence and the appropriate synonym's contextualised embedding in the same sentence (sentence index indicated as n), when using the former to replace the latter in the sentence (e.g., $cos(ball_{s_1}, sphere_{s_1})$). We expect this similarity to increase over layers: while at superficial layers the synonym and the target may be rather different, especially if the polysemous target is used in an infrequent sense, we expect the two representations to grow similar once context exerts its influence, since both words fit in context and are supposed to share their meaning.
- $cos(target_{s_n}, inappropriate_{s_n})$: the cosine similarity between the target's contextualised embedding in a sentence and the inappropriate synonym's

contextualised embedding in the same sentence (sentence index indicated as n), when using the latter to replace the former in the sentence (e.g., $cos(ball_{s_1}, party_{s_1})$). We expect this similarity to increase less than the previous one. It might be tempting to hypothesize that this similarity should decrease over layers, but we know from previous work that BERT's space is strongly anisomorphic (Ethayarajh, 2019), especially at higher layers, so words will have higher similarity values just by virtue of this. Moreover, in our first set of analysis, we saw that contextualisation shifts the representation of a semantically dissimilar replacement towards that of the target. However, for contextualisation to reflect a psychologically plausible process, we contend it is sufficient to observe that *ball* becomes, over layers, more similar to the appropriate synonym *sphere* than to the inappropriate synonym *party* when these words' representations are extracted from the same sentence *The player swung but missed the ball.*

We analyzed the data using Generalized Additive Mixed Models (Baayen et al., 2017, GAMMs). In a first model, we predicted $cos(target_{c1}, target_{c2})$ as a function of layer (included as a simple smooth), including a random intercept over the combination of target and synonyms, which indicates which observations at different layers pertain to the same trial. We expect the simple smooth of the independent variable to have a negative relation with the dependent variable, reflecting the fact that the two representations are identical at the most superficial layer (since the word type is the same) but progressively grow different following the different sentence context.

We then fitted a model to analyze whether the polysemous target grows closer to the appropriate synonym than to the inappropriate synonym. To this end, we fitted a GAMM including layer as a simple smooth, condition [appropriate | inappropriate] as a parametric term, an interaction between the two, and random intercepts for the combination of target and synonyms to group cosine similarities pertaining to the same pair of sentences over layers. To eliminate the possible confound brought by the anismorphic embedding space, we Box-Cox transformed and z-standardized cosine similarities separately for each layer. We hypothesize that the standardized cosine between target and appropriate synonym grows over layers while the standardized cosine involving target and inappropriate synonym decreases. We thus expect to observe a significant interaction between condition and layer.

Comparison to the ITS model. As a point of comparison, we then turned to extract representations using the Instance Theory of Semantics (ITS) model (Jamieson et al., 2018), a psychologically-motivated contextualized representation model (see the respective section in the Introduction). We provide a detailed explanation of the equations behind the

ITS model in the Appendix Mathematical details of the Instance Theory of Semantics (ITS, as well as a Python implementation of the ITS model in the OSF repository[14].

In detail, we implemented two versions of the ITS, changing the underlying corpus. The first reproduces the model tested by Jamieson et al. (2018) in their first presentation of the model. Memory traces come from the passages in the TASA corpus, which are made available on the personal page of the first author together with an R implementation of the model to reproduce results from the paper: we took the input passages as they were pre-processed by Jamieson and colleagues. Each passage is a few hundred words long, with a vocabulary of ∼65,000 word types and a little more than 35,000 passages, each yielding a memory trace. We chose to test the very model which was tested by the original authors as it was shown to provide interesting empirical results for both isolated words and homonyms in minimal contexts. However, since BERT embeddings are derived from a different and larger corpus such as the CoCA, we also implemented a version of the ITS which uses the CoCA to derive traces. Moreover, some of the sentences from the WiC dataset mostly consist of function words: using a corpus from which these have been removed (as in the pre-processing chosen by Jamieson et al. (2018), prevents us from modeling all items in our datasets, introducing a possible bias. We thus implemented a second ITS model using the CoCA and a slightly different pre-processing pipeline. We kept the same dimensionality for the word vectors. We further pre-processed the corpus to exclude strings in all caps, which typically indicate extra-linguistic tags in the CoCA, and punctuation marks. Finally, we lower-cased all strings. We considered each document from the CoCA to be a passage to encode in memory, although a document in the CoCA is considerably longer than a passage in the TASA.[15] We did not try to optimize the model, e.g., changing the pre-processing pipeline or other aspects of the retrieval procedure, just like we did not fine-tune the pre-trained BERT base model, aiming to test whether the general architecture - also instantiated in a way that has had empirical success in the past - meets our desiderata about the relation between context and lexical representations. Finally, we discarded word types

[14] The memory footprint of this model increases considerably with the vocabulary size, as a new high dimensional word vector has to be generated for each word type, and with the number of passages for which a trace is derived: we provide two implementations, one which uses vectorized operations and is faster but also more memory intensive, and another which computes operations sequentially, resulting in slower and less memory intensive computations. We provide the opportunity to speed this up through Python multiprocessing, but the code has only been tested on a Linux server.

[15] There are multiple options as to what to consider as the units to derive traces from: we considered using sentences, which however were too many to compute with and sampling introduced a further degree of freedom we preferred to avoid. Ultimately, deriving a trace for each document in the corpus preserved a thematic unity in each passage that we contend still captures the purpose of the model, although it differs from the original implementation.

with a frequency lower than 50 (but retaining words which feature in the target datasets used in this and following experiment) in the CoCA documents, to have a more manageable vocabulary and exclude word types with a lower frequency of occurrence. Despite these cares, the word *conservatorism* which features in the WiC dataset as a target synonym did not appear in the corpus, forcing us to discard a trial from the analyses. The pre-processed corpus contains close to 250,000 passages (each yielding a memory trace) and a vocabulary of ∼95,000 word types.

We also pre-processed the materials in the WiC dataset following the same approach, removing punctuation, and excluding word types that did not feature in the vocabulary of the model, and lowercasing all tokens. The ITS model derived from the TASA corpus does not cover a few of the words in the items of our benchmark datasets, and as such cannot capture some items, which we will drop in the statistical analysis.

The target measure is the cosine similarity between the echoes of the sentence and the synonym, with two conditions: appropriate - when the synonym matches the correct sense of the polysemous target word in the sentence -, and inappropriate - when the synonym matches the other sense of the polysemous target word in the sentence. We expect the cosine similarity between the echoes of sentences and synonyms to be lower in the inappropriate condition, and tested this hypothesis by fitting a linear mixed model with the cosine similarity as dependent variable, condition (as a binary variable) as the independent variable, and random intercepts for target words.

Results

In the first GAMM predicting the cosine similarity between target word's representations in different sentences extracted using BERT using layer as the independent variable, the predictor has a significant effect (BERT layer: $edf = 8.922$, $Ref.df = 8.998$, $F = 753.7$, $p < .001$), which is negative and wiggles at deeper layers (see Figure 12, left panel). This confirms that the contextualized representations of the polysemous target tend to diverge over layers. This is especially remarkable considering the anisomorphism of the underlying embedding space, which naturally pushes representations closer together.

Results of the second GAMM confirm that the target interaction between layer and condition (a boolean indicating whether the synonym-context combination is appropriate or inappropriate) improves model fit over a simpler model only including the linear combination of layer and condition[16], with a ΔAIC of 107.81 points. Visual inspection of the effects

[16] We compared the target model with a simpler model not including the interaction using the *compareML* function from the *itsadug* R package (van Rij et al., 2022). For the purpose of using this technique, we fitted the GAMMs using Restricted Maximum Likelihood.

(Figure 12, right panel) further shows that the interaction goes in the expected direction, with standardized similarities significantly increasing over layers for appropriate synonyms ($edf = 2.422$, $Ref.df = 3.014$, $F = 18.82$, $p < .001$ and decreasing for inappropriate synonyms (the smooth is the exact mirror since for each congruent pair there is always another incongruent one). We see that the difference grows steadily at initial layers and then stabilizes towards deeper layers.

Turning to the ITS models, the similarity between sentences and synonyms is not significantly different for appropriate and inappropriate synonyms (TASA: $b = -0.0133$, $se = 0.007$, $t = -1.804$, $p = 0.0719$; CoCA: $b = -0.0424$, $se = 0.05554$, $t = -0.766$, $p = 0.444$).

Discussion

This analysis thus shows that contextualisation works in a meaningful way, confirming empirical results from WSD models using BERT (Pilehvar & Camacho-Collados, 2018). Our results further characterize how BERT supports accurate WSD: polysemous words and homonyms have similar representations at superficial layers when used in sentences which trigger different senses, reflecting a stable semantic core. Context integration then operates on this semantic core, pulling the representations away from each other and closer to those of appropriate synonyms for each sense (Cevoli, Watkins, Gao, et al., 2022). This is not merely an epiphenomenon of the structure of the embedding space: appropriate synonyms in the same context are more similar to the target when compared to the inappropriate synonym. Deeper layers are more sensitive to context.

It is important to keep in mind that these pairs are rather noisy in that the polysemous target may be more semantically related to one of the two synonyms than the other. For example, one trial features *place* as the polysemous target, with synonyms *passage* (synset number bn:00062707n in BabelNet) and *space* (synset number bn:00062701n in BabelNet) as synonyms in the sentences *He lost his place on the page.* and *A political system with no place for the less prominent groups.*, respectively. *Space* and *place* are more related than *passage* and *place* when considering static representations to begin with, and yet the representation of *place* in the first sentence grows closer to *passage* than to *place*. The fact that we observe a reliable effect of layer even in the presence of such noise is a further indication of the substantial influence of the contextualisation process on semantic representations.

The observation that the deepest layers exhibit a stable difference between appropriate and inappropriate synonyms is of further interest as it suggests that large differences in meaning that run between separate senses of homonyms and polysemous words

are reflected already at mid layers. These are also the layers where we observed the best fit with explicit ratings of semantic similarity and relatedness in Experiment Experiment 1: Semantic relatedness and similarity norms. Together, these two pieces of evidence suggest that BERT recognizes coarse differences in meaning before getting to the deepest layers, confirming previous evidence about the specialization of different attention blocks for different linguistic phenomena (Rogers et al., 2020).

Finally, the analysis of similarities derived using the ITS model contrasts with the evidence provided by Jamieson et al. (2018) that the ITS model can account for meaning modulations that cross sense boundaries. The failure of the model trained on both corpora suggests, thus, that the model may not very robust. Moreover, the presence of stop words may be particularly detrimental to this model: since these tend to frequently occur in all passages, they may overshadow subtle differences in the distribution of content words. We considered the distinctiveness of each passage by considering the Jaccard coefficient between the word types in each passage from the TASA corpus and then each passage of the CoCA (both after pre-processing): while the set overlap is around 0.05 for the TASA, it is around 0.3 for the CoCA, confirming that passages tend to overlap a lot more in the CoCA in terms of lexical items. This is however a characteristic of the input, and we contend that a successful model should be able to discard useless or misleading information: the weighting foreseen by the ITS does not seem to handle the noise in memory traces sufficiently well for the model to handle realistic sentences.

In this study, we considered the relation between fully contextualised representations. In the next one, instead, we consider how contextualised representations of homonyms and polysemous words in disambiguating contexts interact with the corresponding prototype representations.

Experiment 6: Priming in context

In the previous study, we have seen that contextualised BERT representations are able to successfully disambiguate word senses in sentence contexts, thereby focusing on *context-dependent representations* alone: we only compared contextualised representations, considering the polysemous word and two sense-related synonyms. Here, we address another crucial issue in contextualisation: What is the relation between context-dependent representations of ambiguous word types and prototype-representations, which inevitably collapse different senses? In order to answer this question, we rely on the stimuli from the study by Till et al. (1988) (see also Kintsch, 1988). We follow the approach outlined in M. N. Jones et al. (2006), analysing the relations across stimuli and comparing the patterns of similarity produced by the model with the qualitative pattern observed in the experiment.

Data

The experimental material consists of 56 texts from the original study by Till et al. (1988, Appendix A). These were written around 28 ambiguous words, with two sentences per ambiguous word (such as *"For the third time, the worried player swung but missed the ball. He knew what the coach would say."* versus *"The lonely maiden had great hopes as she dressed to go to the ball. This was her only opportunity to meet people."*. As for these examples, the texts always consisted of two sentences. In 26 texts, the critical ambiguous word was presented as the last word of the first sentence (as in the example provided here). In the other 30 texts, it was presented as the last word of the second sentence instead.

Each text was paired with a context-appropriate associate word for the critical ambiguous word (*bat* for the first text, *dance* for the second), taken from the dataset by Cramer (1970). These associate words served as the real-word targets for the lexical decision task: After presentation of the critical, ambiguous word (here, *ball*), participants were required to make a lexical decision task for a target word which was either a pseudoword or, critically, a context-appropriate associate word (*bat* for the first sentence, *dance* for the second) or context-inappropriate associate word (*dance* for the first sentence, *bat* for the second)[17]. The context-appropriate word for one sentence in the text pair was always used as the context-inappropriate word for the other text.

In their Experiment 1, Till et al., 1988 found no response time differences between the appropriate and inappropriate associate when it was presented very shortly after the critical ambiguous word (stimulus onset asynchrony (SOA) = 333 ms), but observed significantly faster responses for the context-appropriate associate after some delay (SOA = 1000 ms). In a second experiment, Till et al. (1988) found the difference also at short SOAs, but still smaller than at long SOAs, confirming the main pattern of Experiment 1. This suggests that the sentence contexts activate a disambiguated meaning representation of the critical (originally ambiguous) word, which in turn primes context-appropriate words to a stronger degree than context-inappropriate ones (this disambiguation process appears to require some time to unfold, as indicated by the effects of SOA).

Our final dataset consists of 28 critical words (e.g., *ball*) used in two different sentences in two different senses (see the description above), each paired with two target words (e.g., *bat* and *dance*) each associated to either of the senses in which the critical word is used, for a total of 56 different target words, on which lexical decision RTs were collected

[17] The original study by Till et al. (1988) also included additional experimental conditions with appropriate or inappropriate inferences, which we did not consider for the present analysis.

in the original study. Each item (critical and target words) was embedded at the usual 14 Bert layers, for a total of 1568 data points.

Methods

In this experiment we compare the contextualised representation of the critical polysemous word (which appears in a sentence which is supposed to affect its meaning) to the prototype representation of the target word (presented in isolation). With respect to the studies presented in Replacements, the current experiment provides a complementary test of whether models like BERT can be fruitfully used to compare contextualised and non-contextualised representations: Priming effects in context are one case in which the possibility to do just this can help us understand semantic effects better.

Our hypothesis is straightforward: Extensive previous research has shown that cosine similarities between distributional semantic vectors are well-suited to capture priming effects in lexical decision (Gatti et al., 2022; Günther et al., 2016b; M. N. Jones et al., 2006; Mandera et al., 2017; Petilli et al., 2021). If contextualised BERT representations are cognitively adequate token-level meaning representations for words in sentence contexts, then we have to expect that the cosine similarities between the contextualised BERT representations for the critical words and the prototype representations of the (context-appropriate and context-inappropriate) target words in a priming LD experiment as the one by Till et al. (1988) exhibits a pattern consistent with behavioral evidence from Till et al. (1988) (in line with the methodology in M. N. Jones et al., 2006).

We fed the stimulus materials (sentence prompt, critical word, and target word) to our BERT embedder, to obtain the following three representations:

- $critical_c$: the contextualised embedding of the critical polysemous word (*ball*, in the example above) in the context of the sentence, at the usual 14 layers;
- $target_{non-cont}$: the non-contextualised prototype embedding of the target word (*bat* or *dance* depending on the sentence), which participants had to rate as a word or pseudoword, extracted at the usual 14 layers;
- $target_{avg}$: the averaged prototype embedding of the target word (*bat* or *dance* depending on the sentence), which participants had to rate as a word or pseudoword, extracted at the usual 14 layers

Note that, unlike in the WiC experiment, the embeddings of the target words (*bat* or *dance*) do not change when they are used as appropriate or inappropriate targets, since the representation is a prototype. Appropriateness or inappropriateness of the synonym here

does not depend on a difference in the representation of the target, but on a difference in the embedding of the critical word, which is the same in both conditions but embedded in different sentences. From these representations, we computed the following cosine similarity scores in the appropriate and inappropriate condition:

- $cos(critical_c, target_{non-cont})$
- $cos(critical_c, target_{avg})$

to capture to what extent the polysemous critical word is similar to the prototype representation (non-contextualised or averaged) of the target word, which was judged for lexicality in isolation. The hypothesis is that the similarity grows in the appropriate condition but decreases in the inappropriate condition (once the anisomorphism of the underlying embedding space has been accounted for by standardising similarities).

Turning to the ITS models, we derived echoes using the same memory traces derived in Experiment 5: Word Sense Disambiguation. The target measure is the cosine similarity between the echoes of the sentence (including the polysemous critical word) and the target word which elicited the lexical decision, with two conditions: appropriate - when the target word matches the sense of the polysemous critical word in the sentence -, and inappropriate - when the target word matches the other sense of the polysemous critical word in the sentence. Much like in Experiment 5: Word Sense Disambiguation, we expect the cosine similarity between the echoes of sentences and target words to be lower in the inappropriate condition, and tested this hypothesis by fitting a linear mixed model with the cosine similarity as dependent variable, condition (as a binary variable) as the independent variable, and random intercepts for critical words.

When analysing BERT representations, we fitted a GAMM for each prototyping method, including condition [appropriate | inappropriate] as a parametric term, layer as a simple smooth, an interaction between the two, and random intercepts for each combination of critical and target word. As usual, to avoid that the underlying anisomorphism of the embedding space at deeper layer confounds results, we Box-Cox transformed and z-standardized the cosine similarities by layer.

Results

For both prototyping methods we see that the target interaction improves model fit over a model only including a simple linear combination between layer and condition ($\Delta AIC = 112.43$ for averaged prototypes; $\Delta AIC = 53.79$ for non-contextualised prototypes). Visual inspection of the effects (see Figure 13, left panel: averaged prototypes; right panel: non-contextualised prototypes) confirms that the interaction goes in the expected direction,

with appropriate target words given the sense of the critical word in context growing more similar over layers (averaging: $edf = 2.399$, $Ref.df = 2.986$, $F = 20.23$, $p < .001$; non-contextualised: $edf = 3.197$, $Ref.df = 3.967$, $F = 7.736$, $p < .001$). The cosine similarity between the critical word and the inappropriate target word, on the contrary, decreases over layers (the smooth terms for layer and the inappropriate condition are exact mirrors of those reported above since each critical word occurs in two sentences), showing that context operates in the expected direction. Importantly, at layer -1 (i.e., for the static embedding) the cosine similarity is not different between the two conditions. We also see that the statistical models explain a substantial amount of deviance (63.9% for averaged prototypes and 52.8% for non-contextualised prototypes), confirming that the interplay of contextualisation and condition accounts for much of the phenomenon at hand. Comparing prototyping methods, we again see that non-contextualised prototypes become unreliable at very deep layers, where the presentation of an isolated word to a model trained on rich sentence context creates instability in the representations. With averaged prototypes, on the contrary, the effect is stronger moving from superficial to mid layers and again plateaus at deeper layers.

Turning to the ITS models, the similarity between sentences and synonyms is not significantly lower in the inappropriate condition when using both corpora to derive traces (TASA: $b = -0.0134$, $se = 0.0145$, $t = -0.920$, $p = 0.361$; CoCA: $b = -0.0081$, $se = 0.1116$, $t = 0.073$, $p = 0.942$). Moreover, the similarity between target words and sentences including ambiguous critical words is not significantly different from 0 also in the appropriate condition, with a strongly bimodal distribution and the two modes around 0.8 and -0.8 in cosine similarities for both instantations of the ITS model we probed. This distribution suggests that, next to missing the distinction between appropriate and inappropriate target words given the ambiguous critical word embedded in a sentence, the resulting echoes do not capture semantic relations in general. This result stands in contrast with evidence provided by Jamieson et al. (2018); we address this in the discussion.

Discussion

In this experiment, we observe a similar pattern to the one reported for word sense disambiguation using the WiC dataset. The two analyses share many traits, but whereas the analysis of the *WiC* dataset compared fully contextualised representations, the current analysis extends it to comparing contextualised and prototype representations. In line with evidence from studies discussed in the Replacements section, this further confirms that BERT is a viable tool to investigate the interplay between words in and outside of context.

Context operates in a systematic way on representations and shifts their position in the space relative to other words, as evidenced by the analysis we just described.

The pattern we show mirrors the findings reported by Till et al. (1988): At superficial layers we observe no or only scarce differences between appropriate and inappropriate target words, which grow larger at deeper layers. By drawing a parallel between SOA and degree of contextualisation (with the simplified assumption that a context-sensitive processing mechanism unfolding over time, as in Kintsch, 1988's model, is roughly mirrored by the passing of information through BERT's layers), the lack of differences at superficial layers reflects the lack of differences in RTs at short SOAs, whereas the differences we observe at deeper layers can be taken to mirror the difference in RTs at long SOAs. In psychological terms (Kintsch, 1988, 2001), at short SOAs, language processing has not yet managed to integrate contextual information and thus the decision operates on representations which still heavily depend on the semantic core, which pertains to the type more than to the context. At longer SOAs, however, when context has been integrated and the representation has been shifted towards a specific sense which either fits or does not fit with the semantics of the target, RTs are faster for appropriate targets, just like the representations of the critical word and the appropriate target have grown closer. Of course, the underlying assumption of a parallel between processing time and contextualisation unfolding through BERT's layers has to be empirically investigated in future studies.

In addition to mirroring behavioral results obtained in a seminal study on semantic processing of ambiguous words, our analysis further shows that contextualisation in BERT produces contextualised representations whose similarity to prototype representations changes in semantically meaningful ways. We discussed in the Introduction how meaning variation could be recovered post-hoc from static embeddings (Beekhuizen et al., 2021), suggesting that usage patterns provide models with sufficient information to learn about different senses. BERT appears to implement learning mechanisms which equip it with the ability to magnify certain aspects of the static representation and suppress others depending on context. This process results in different similarity profiles for the same word type in different contexts.

Unlike what we observe with BERT embeddings, the echoes derived using both instantiations of the ITS model did not capture the hypothesized relations, and showed other concerning properties. Differently from the setup used by Jamieson et al. (2018) for their evaluation of joint probes to capture semantic ambiguity, the probes used in this experiment included much longer sentences. This poses a problem for the ITS model as currently specified, since when probes contain multiple words the activation used to weigh traces when computing the echo is multiplied several times. This yields very small echoes, that tend to

shrink the more elements a probe contains. Smaller activations result in smaller traces, which result in smaller echoes: this process, we suggest, results in very brittle echoes which may look more or less similar to the echoes of other probes following noise more than signal from semantic relations found in the input. Crucially, this behavior does not only manifest in the model derived from the CoCA but also in that derived from the very same passages which yielded a successful model in a previous study by Jamieson et al. (2018). Moreover, whereas when considering BERT the analysis of the WiC dataset and the current dataset presented important conceptual differences (a comparison of contextualised embeddings in the WiC dataset against a comparison of contextualised and prototype embeddings in the current experiment), the two analyses using the ITS model are isomorphic: the echo of an isolated word is compared to the echo of a complex probe (a sentence), overshadowing important differences. As we have already mentioned, the ITS cannot derive the echo of a word in context but only the echo of the context. Given the empirical results and these considerations, we suggest that the representations produced by the ITS model as currently implemented do not offer a particularly useful tool to study meaning modulations in natural language. Conversely, our analyses do suggest that contextualised embeddings derived using transformer models such as BERT do offer a fruitful tool to study meaning representations in isolation and in context.

Experiment 7: Meaning Modulations

The previous two studies investigated whether BERT captures differences in meaning for homonyms and polysemous words. However, as argued in the Introduction to this article, this reflects only a (prominent) portion of the much wider phenomenon of contextualized, token-level meaning representations. In the present experiment, we therefore zoom in on subtler context-dependent meaning modulations that, we contend, exist even when such modulations may not result in discrete meanings (i.e., different senses). We operationalize this intuition as different associations being elicited by a word used in the same sense in different sentence contexts. To expand on the *tomato* example we mentioned in the Introduction, we can use the word *tomato* to refer to the exact same referent, but by using it in different sentences, the same word indexing the same referent may evoke different associations (Zwaan et al., 2002). Saying that *I plucked a fresh tomato from the garden.* may evoke associations with plants, a round shape, and solid texture, while saying *Parts of the tomato are still visible in the sauce.* may evoke associations with spaghetti, hot temperatures, and mushy texture. We argue that a good model of semantic memory should be sensitive to these modulations, and thus the representation of *tomato* in context should shift systematically as a function of context, not only when context differs between different

senses of a word but also following these subtler meaning modulations. This analysis is the first empirical study of how lexical representations of words used in different sentence contexts yet still in the same sense vary systematically with respect to prototype representations of semantically related lexical items. We argue that such an analysis can offer a more detailed characterization of how meaning shifts as a function of abstract prototype representations and linguistic context, an interplay that has so far only be addressed for words used in different senses.

Data

In order to explore this issue, we created the *Meaning Modulation* (MeMo) dataset, consisting of 49 target words, each used in at least three different sentences in the same sense but with subtle meaning modulations. For each target we manually chose an associated word that reflects the meaning of the target in the sentence. For example, consider the target word *happy* (see Table 2), defined as *Emotion experienced when in a state of well-being* in WordNet 3.0 or as *Emotional state characterized by feelings of enjoyment, pleasure, and satisfaction.* in Wikipedia. *MeMo* contains three sentences featuring *happy* as target word, each matched with a context-appropriate associated word (see 2). The associated words of the *other* sentences containing the same target are taken as context-inappropriate associates. The dataset contains nouns, verbs, and adjectives as both target and associated words. We generated items based on our own intuitions and finalised the associated words by consensus among all five authors. The sample size was determined based on our own intuition as we did not have prior evidence on which to base a power analysis, nor we had a precise expectation of the smallest effect of interest nor of the variance-covariance matrix involving random effects (Lakens, 2022). With 49 target words, a minimum of three associates per target word, and considering the 14 layers in BERT base, there are in total 7,140 observations in the MeMo dataset, with 510 unique combinations of target word, associate, and sentence. 2,184 observations pertain to appropriate associates (156 clusters, with 14 observations each, one per layer) and 4,956 observations pertain to inappropriate associates (354 clusters, with 14 observations each, one per layer). We performed simulations to get a sense of how robust the effect is and whether we find evidence the current dataset may lead us to report an effect which is not there. These simulations are reported in Appendix MeMo dataset reliability analysis.

Our hypothesis is that the contextualised representation of the target word derived from a sentence will shift towards the prototype representation of the appropriate associated word, and further away from the inappropriate associated words. Importantly, *all* associated words have some considerable degree of association with and similarity to the target word.

Target	Sentence	Associate
happy	My grandparents married in 1949 and had a happy marriage.	lasting
happy	The owner was so happy to see his dog return home after it had been missing for three days.	relieved
happy	I'm always happy to discover a new little restaurant.	excited

Table 2

Example items from the MeMo (Meaning Modulations) dataset. There are 49 target words in the dataset, each used in at least three sentences and paired to an associated word.

In this, the *MeMo* dataset is particularly challenging. Unlike the previous two analyses, we go beyond discrete senses in this experiment and continue to explore whether contextualised and prototype representations can be fruitfully compared and are shown to encode appropriate semantic relations.

Methods

We derived the same representations outlined in Experiment 6: Priming in context when describing the modeling study of the stimuli from Till et al. (1988) and used them to compute the same cosine similarity scores between the word in the sentence and the prototype representations of the associates, except that in this case the inappropriate associates are more than one. We analyze the cosine similarities taking the same approach outlined in Experiment 5: Word Sense Disambiguation and Experiment 6: Priming in context and formulate a consistent hypothesis: We expect the similarity between contextualised target word and appropriate associate to grow over layers, while the similarity between the contextualised target word and the inappropriate associates is expected to decrease over layers (always once the anisomorphism of the underlying space has been accounted for). We thus expect a significant interaction between layer and condition.

Turning to the ITS models, we again used the same two memories we used in the previous experiments, one derived from the TASA corpus and one derived from the CoCA. The target measure in the current experiment is the cosine similarity between the echoes of the sentence (including the target word whose meaning modulations we aim to capture) and the associate word, with the usual two conditions: appropriate - when the associate aligns with the meaning modulation afforded by the sentence, and inappropriate - when the associate does not align with the meaning modulation afforded by the sentence. As for both previous datasets, we expect the cosine similarity between the echoes of sentences and associates to be lower in the inappropriate condition, and tested this hypothesis by fitting a

linear mixed model with the cosine similarity as dependent variable, condition (as a binary variable) as the independent variable, and random intercepts for target words.

Results

We do indeed observe a significant interaction between condition and layer for averaged BERT prototypes ($\Delta AIC = -60.40$ with respect to a simpler model using the linear combination between layer and condition), but not for non-contextualised prototypes. We thus limit ourselves to the analysis of averaged prototypes (which, as already established across previous studies, in general appear to be better suited to capture semantic relations involving prototypes). Visual inspection of the effect (Figure 14) shows that layer has a positive and rather linear effect on standardized cosine similarities between target word and appropriate associates ($edf = 1.616$, $Ref.df = 2.004$, $F = 21.71$, $p < .001$) whereas it has an equally linear but negative effect for inappropriate associates ($edf = 1.027$, $Ref.df = 1.054$, $F = 17.95$, $p < .001$).

As observed in the previous experiment, the difference between appropriate and inappropriate associates was not predicted to be significantly different with either of the ITS models (TASA: $b = 0.0054$, $se = 0.0085$, $t = 0.638$, $p = 0.524$; CoCA: $b = 0.0425$, $se = 0.0661$, $t = 0.643$, $p = 0.52$). Moreover, the similarity between sentences and appropriate associates was not different from 0, with a bimodal distribution that mirrors the one described for the previous experiment.

Discussion

The pattern observed here thus reflects the one found when analysing stimulus materials from Till et al. (1988). However, beyond this previous observation, the present results show that BERT representations shift in context following intuitions about associations in meaning that are not simply confined to different related or unrelated senses, but also extend to meaning modulations within the same sense of a word. This opens the possibility to study meaning differences in context to a more fine-grained level and better investigate how context affects meaning representations with respect to context-independent ones.

It is interesting to note that the effect size of layer shrunk for *MeMo* (and disappeared altogether when considering non-contextualised prototypes): the difference between appropriate and inappropriate words to the target in context is arguably larger when the target is used in different senses (Cevoli, Watkins, Gao, et al., 2022), as opposed to when it is used in the same sense but with subtler meaning modulations (as is the case in *MeMo*). The instability of non-contextualised prototypes, especially at deeper layers, washes

out such a difference; however, the effect persists when deriving prototype embeddings by averaging representations.

The shape of the effect is also interesting, albeit only at a speculative level: while we saw curves plateauing for both WiC (Pilehvar & Camacho-Collados, 2018) and the stimuli by Till et al. (1988) – suggesting that the very deep layers did not exert a further influence on the target representations – we observe a linear effect for *MeMo*. This sits well with the focus of the stimuli in this dataset on finer modulations in meaning that are virtually entirely dependent on context. Our results thus suggest that BERT can even capture such fine-grained modulations, and it does so more effectively the more contextualized its representations are.

Finally, the ITS models did not exhibit the expected behavior. The MeMo dataset contains fairly long and complex probes, rather different than the ones used in the WiC dataset and in the study by Jamieson et al. (2018) but more in line with those used in the stimuli of Till et al. (1988). The ITS model as currently specified does not seem able to handle the complexity of such probes, regardless of the presence of very frequent items like stop words.

Meaning in Context: Summary

The last set of analyses, together with previous ones, satisfies all the desiderata we put forward at the beginning of this paper. Representations have a stable semantic core which enters meaningful relations with contextualised representations. Moreover, words cannot simply be exchanged with one another while leaving it to the context to disambiguate the meaning: replacing inappropriate synonyms in context results in lower similarity to a polysemous target. However, context does influence the representation of a word token, shifting it with respect to the static semantic space and other contextualised representations affected by the same context.

Importantly, we showed that BERT's representations capture meaning shifts both across and within word senses, and do so at different places in the architectures.
In Experiment 5: Word Sense Disambiguation and Experiment 6: Priming in context, we see that representations of appropriate and inappropriate associate words change steadily over the first few layers but the difference stabilizes around layer 9, suggesting that differences across sense boundaries are captured early. On the contrary, meaning modulations within a same sense, targeted in Experiment 7: Meaning Modulations, are captured also at deeper layers, where the influence of context is more pronounced (as expected). Recent work by Trott and Bergen (2023) highlights how humans are influenced by differences in discrete word sense as well as by continuous meaning variations, and also that BERT fails to capture

this tension. Our evidence suggests that the model might, provided that one considers how representations shift throughout the architecture's latent space and how context is integrated, althouth a thorough empirical evaluation is necessary to assess this hypothesis.

Next to showing that BERT embeddings reflect lexical ambiguity along discrete senses, we provide the first piece of evidence that the model can account for subtler modulations in meaning, operationalised as context-dependent associations of unambiguous words used in different contexts. The MeMo dataset offers a first test bed for probing models on their ability to reflect the intuition that meaning can change within the boundaries of a same conventional word sense, slightly but consistently and *predictably* altering the pattern of associations the word elicits in a specific context as opposed to a different context or in isolation. More than anything, this set of studies further characterizes meaning as particularly fluid and shows that a model such as BERT can capture this fluidity at various scales and considering both exemplars and prototypes (Murphy, 2002b).

Finally, while we cannot exclude that an optimized version of the ITS model would fit the data well, we do contend that the problem lies in the weighting scheme put forward in this model rather than in a lack of optimization - also considering that we evaluated the very same model which was shown to have empirical success on context-dependent meaning variations across sense boundaries. Since echoes result from the weighted sum of all memory traces and since the weighting depends on the similarity between each item in the probe and the memory trace, the lack or presence of similarity between any item in the probe and the trace can make the resulting echo rather brittle: It is unlikely that in a complex enough sentence all word types contribute to the meaning of the probe in the same way by virtue of comparable similarities with all traces. On the contrary, a model that implements a learning strategy grounded in context-dependent prediction and updates based on errors, such as transformer models, appears better equipped to downplay frequent yet uninformative elements than a weighted analogy function based on similarity over recordings of experience. Furthermore, the attention mechanism implemented by BERT and similar models aims precisely to learn which other words in the context matter most to shift the representation of a certain lexical element. Finally, as already pointed out in the Introduction, the ITS model inherently collapses the meaning of a word in context with the meaning of the context, a confusion that obfuscates the contribution of individual words to the context and prevents an analysis of how each word interacts with its linguistic context.

General Discussion

In this work we analyzed to what extent the semantic representations learned by recent transformer-based language models, here exemplified by BERT, can offer a useful tool to

investigate psycholinguistic effects pertaining to semantics. Crucially, we used the same model to derive context-dependent and context-independent representations to investigate the tension between exemplar-based (implemented in our model as the token-level representations) and prototype-based (implemented as type-level static representations) accounts of semantic representation (Murphy, 2002b). First, we established that the prototype embeddings derived from BERT capture behavioral patterns (semantic similarity and relatedness, Bruni, Tran, et al., 2014; Hill et al., 2015; semantic richness effects in lexical decision, Yap et al., 2011; and prime-target similarity in primed lexical decision, Hutchison, Balota, Neely, Cortese, Cohen-Shikora, Tse, et al., 2013; Koriat, 1981), attesting to their viability as implementations of a semantic memory model. We then showed that the contextualized embeddings learned by BERT have a stable semantic core, and that the effect of the sentence context on this representation interacts with said semantic core: Replacing words in context yields representations of the replacements that shift closer to the replaced word (as evidenced by the similarity of the lexical representations as well as of their neighborhoods). Moreover, this shift is modulated by the similarity of the static representations in a psychological model of semantics (De Deyne et al., 2019). Finally, we show that contextualisation follows meaning variations at the level of discrete senses (Cevoli, Watkins, Gao, et al., 2022; Li & Joanisse, 2021; Rodd, 2020), but also captures subtler yet systematic variations pertaining to words used in the same sense in different contexts. The simulations we report thus allow us to conclude that BERT embeddings meet our desiderata for psychologically valid semantic representations: words do have a semantic core, which captures behavioral patterns observed for isolated words; they cannot be exchanged randomly in context, but carry a distinctive footprint which originates in their co-occurrence patterns; and they are sensitive to context, yielding meaning variations which exist on a continuum.

Our study highlights several psycholinguistically relevant properties of these embeddings, which are discussed in more detail in what follows. First, we showed that explicit and implicit semantic effects are best captured at intermediate layers especially when deriving a word's prototype representation by averaging a handful of contextualised representations for that word. These averaged prototypes provide a consistently better fit than the de-contextualized prototype representations BERT learns and stores, highlighting how token-level representations can complement type-level ones and in relation with specific aspects of the model architecture. This extrapolation of prototypes from exemplar representations is in line with the fundamentals of psychological prototype models (Rosch, 1978) as well as recent computational work on exemplar and prototype representations (Battleday et al., 2020; Günther et al., 2023). Moreover, we observed that the deepest layers struggle with dealing with isolated words and that the very last layer is biased by context

over a word's semantic core, whereas intermediate layers capture both prototype effects and interactions with context-dependent representations more reliably. This highlights structural properties of these models and the representations they learn that should inform as well as constrain their application to study psycholinguistic effects. Finally, our study is the first to show how meaning variation systematically affects individual occurrences of each word whether they encode different senses or not, and that these graded difference is captured by the architecture (which, as we discussed before, builds upon but significantly extends psychological semantic representation models such as by Rodd, 2020). Whereas differences between tokens of different word senses tend to diverge already at superficial layers, reflecting a strong interaction between context and semantic core, differences between tokens of a same sense keep growing until the very last layers, underlying how crucial context is in modulating meaning variations in the face of meaning stability. Contextualised embeddings can thus be fruitfully used to study these modulations further and how we encode and process them, contributing to the debate on exemplar vs. prototype accounts of semantics by offering a possible balance point: a system that abstracts what can be abstracted from the environment because of its invariance and learns a model of how context can affect these representations so that they can be used flexibly and productively to capture the endless variability of experience.

Implications of the model architecture

When evaluating the BERT representation system from a cognitive perspective, it is useful to go back to its architecture, especially the role of transformer encoders and of the learning objective. BERT (Devlin et al., 2019a) is trained to predict masked words from context as well as to predict the next sentence. Hence, the representations it develops support an accurate execution of these tasks. In particular, the learning objective of word prediction is taken to occupy an important role in language processing (Federmeier, 2007; Huettig, 2015; Willems et al., 2016) and is conceptually similar to that of countless other DSMs (Bojanowski et al., 2017; Li and Joanisse, 2021; Mandera et al., 2017; Mikolov, Chen, et al., 2013; see Lenci et al., 2022). While the focus on prediction in language has traditionally been on comprehension and processing (Frank & Willems, 2017; Monsalve et al., 2012; Willems et al., 2016), recent studies have started to consider the role of prediction in learning (Fazekas et al., 2020; Vujović et al., 2021), suggesting that it provides a productive explanation of behavior. BERT, and similar models, offer a possible account of how learning and processing are inherently linked, with representations being learned and updated as a result of prediction in context.

At a fundamental level, the model will develop similar representations when it needs to output the same prediction, that is the same token. However, because of ambiguity, many tokens in the sentence context will end up supporting very different predictions: observing the word *ball* may nudge the model to predict *bat* but also *gown*. However, the co-occurrence patterns across words are also predictable and the model can learn that, given a certain sentence context, it should prioritize certain words over others because their co-presence is informative with respect to a certain prediction instead of another. For example, the presence of the word *striker* together with *ball* would call for predicting *bat* and not *gown*. This is what attention heads in transformer encoders do: they are trained to attend to context to improve prediction in context. Therefore, when a word is replaced in context, the replacement will be represented slightly closer to the intended target than it would be in a semantically inappropriate context: The model has learned to attend to context and that a certain context calls for a certain prediction, which however has to be combined with what is eventually observed. If the context grows, as in the story of the amorous peanut (Nieuwland & Van Berkum, 2006), the evidence will push the model to substantially alter the representation, following what it learned about the co-occurrence of context and target words. This does not rule out the existence of static, context-independent representations. Rather it highlights how such context-independent representations constrain and influence the accommodation of semantic violations. In the limit, if no co-occurring word has any context-independent representations, then context can mean anything and becomes unconstrained. The *peanut* can be *in love* because the co-occurring words have a certain meaning, which does not fit with *peanuts* but is coherent and informative enough to warp the representation of the target word and derive a representation that fits in context. Our analyses predict that this process can be easier or harder depending on the similarity of context-independent representations of target and context.

These properties of the BERT representation system provide important insights about how human speakers can acquire and develop useful semantic representations that are both sufficiently stable and sufficiently flexible to be used in actual communication (Pickering & Garrod, 2004). In cognitive science, "standard" distributional semantic models have been established as plausible and high-performing learning models for semantic representations (Günther et al., 2019; Hollis, 2017; Mandera et al., 2017), building on the fundamental assumption that humans utilize statistical regularities – that is, context information – in their environment to assign meaning to stimuli (Griffiths et al., 2007; Lenci, 2008). The results of the present study further extend this without changing this underlying assumption (Lenci et al., 2022): After all, BERT also derives its representations by considering context information, it just incorporates a wider notion of context (crossing the

arbitrary sentence boundary incorporated by many previous models) and more elaborate learning techniques. Thus, the model and results of the present study provide a proof of principle of how humans can acquire context-sensitive semantic representations by learning from the statistical patterns in their linguistic experience.

Contextualised representations in psycholinguistics research

Our results on context-dependent meaning patterns highlight that BERT provides an effective implementation of a model of semantics which is aware of the linguistic context in which a word occurs (Cevoli, Watkins, Gao, et al., 2022; Li & Joanisse, 2021; Rodd, 2020). The layered architecture also offers the possibility to study how context is integrated in the lexical representation of a word, and whether this speaks to how humans achieve the same broad goal. This is an important difference with respect to the implementation by Li and Joanisse (2021) which relied on an LSTM network. There, one could consider the static representation of the word on the one hand, and the representation the model develops considering the preceding context and the word itself on the other hand. In this sense, it is unclear whether the context-dependent representation really encodes the context-dependent meaning of the last word (i.e., the word in question) or rather the meaning of the sentence up to that point. The model we consider, on the contrary, allows to more effectively tease the two apart. Moreover, unlike the model by Jamieson et al. (2018), which relies on static representations (encoded using random vectors matched to each word type) and sees context effect arising at the level of the full word combination rather than the individual word, BERT embeddings actually shift in semantic space by virtue of the influence that the sentence context exerts on a learned semantic representations which captures isolated word effects - a feature that random vectors would not have by design.

BERT's semantic representations are more reminiscent of Rodd (2020)'s proposal, although the underlying model does not foresee explicit attractors. Nonetheless, our study on replacements, especially considering neighborhood similarity, hints to the fact that such attractors are an emergent phenomenon in BERT's semantic space, even if no explicit role for them is hard-coded into the model. We saw that contextual representations are dragged towards or away from coherent semantic neighborhoods which reflect stable configurations of the underlying semantic space. Thus, attractors may indeed signal stable states of the model, which correspond to representations that come to occupy coherent positions in the resulting embedding space. However, these attractors are not discrete nor are they all-or-nothing: representations can, and do, exist outside of these, highlighting how meaning variations are best construed as continuous.

Meaning ambiguity beyond word senses

Further challenging the notion of polysemy as an explanatory construct, we showed that BERT is also able to encode differences in word meaning that do not pertain to different senses (Haber & Poesio, 2024). BERT embeddings themselves offer evidence that meaning may be systematically altered by context beyond discrete senses. This further calls into questions the traditional categorical distinction between homonyms, polysemous, and unambiguous words: we observe systematic meaning modulations also for words supposedly used in the same sense. The continuous, distributed, context-dependent representations that models such as BERT offer thus provide a new tool to evaluate semantic aspects in psycholinguistic tasks, which moves away both from the rigid partitioning of thesauri and from the monolithic representations of static DSMs towards a continuous and graded account of the meanings related to a word.

This study further qualifies extant models of how people deal with ambiguity in natural language, which have focused primarily on modeling discrete senses. Rather than aiming to encode or represent different senses, BERT's ability to develop representations that reflect lexical ambiguity (Apidianaki, 2023; Cevoli, Watkins, Gao, et al., 2022; Haber & Poesio, 2024) does not stem from the explicit goal of keeping track of ambiguity, but rather of optimising prediction (Kveraga et al., 2007; N. J. Smith & Levy, 2013). That we can recover clusters of representations which align with word senses is a by-product of what the model learns to do: predicting masked words while exploiting linguistic context to do this better. It does not rely on explicit attractors (contrary to Rodd, 2020's account), but rather achieves this by implementing the basic principle that no two representations are alike unless their contexts of occurrence match entirely. All the model learns are token-level representations: we argue that what we refer to as *polysemy* is a simplified description of the fact that these token-level representations cluster more or less depending on context (Cevoli, Watkins, Gao, et al., 2022).

However, recent research by Trott and Bergen (2023) highlights how people are sensitive to sense boundaries in a way which a model like BERT fails to entirely capture, describing semantic representations as both discrete (with marked differences along word sense boundaries) and continuous (since graded differences in similarity explain more variance than discrete senses). We used the same architecture and focused our attention on graded relations both *across* and *within* word senses, and our results suggest that BERT might handle meaning variations that cross or do not cross sense boundaries in different places of its architecture. In detail, we noticed that two different senses of a homonym do not change throughout the last layers, with the model shifting the representation enough over the first layers. On the contrary, meaning modulations within the same sense are

reflected in subtler but continuous changes in the underlying distributed representation until the very last layer. It might therefore be useful to consider how representation unfolds throughout the architecture to better grasp the discrete aspects of word meaning in this model, next to the graded semantic relations we have focused on in our experiments.

Token-level, exemplar, and prototype representations

The ability to capture fine-grained ambiguities would seem to put BERT closer to exemplar than prototype accounts of semantic representation (Ambridge, 2020; Murphy, 2002a), since each token is encoded differently. However, unlike exemplar-based accounts, BERT does not store each representation. To the contrary, it actively learns a static representation – the embedding at layer 0 – before positional information is integrated. Crucially, the reliability of these representations evidenced throughout our experiments shows that BERT representations encode *meaning dominance* (Rodd, 2020, and references therein), the observation that when presented in isolation, people interpret words according to their most frequent meaning. The relation between contextualised and static representations in the replacement experiment, as well as the patterns observed when modeling Till et al. (1988)'s material and the MeMo dataset, confirm that prototype representations are sensitive to these frequency patterns in natural language and show that BERT prototype do capture meaningful invariances in language experience, since we find stable relations that mirror those found in a psychological model of meaning grounded in free associates (De Deyne et al., 2019). Next to this, however, BERT also encodes a model of how context interacts with these static representations, relying on the many systematic patterns in language to build a model of context influence that operates on stored abstract prototype representations to yield ephemeral token-level representations. We contend that BERT implements a more parsimonious account than storage-based exemplar models and a more flexible one than prototype models, which strikes a productive balance between capturing the invariance in experience together with its idiosyncracies. Crucially, it achieves this by *learning* an abstract model of how the latter affects the former to yield infinitely varied yet systematic representations.

An important comparison in our study involves the way in which prototype representations (i.e., static embeddings) should be derived (Apidianaki, 2023). Next to assessing the viability of embeddings at layer 0 before including positional encoding, we tested two methods which have been widely used in the literature (Bommasani et al., 2020; Lenci et al., 2022): on the one hand, we derived prototype representations by averaging contextual representations extracted from a sample of sentences featuring the target word (*averaged*); on the other, we fed the target word outside of any sentence context to the model

(*non-contextual*). While both approaches show some effectiveness, the former is unequivocally better in that it explains more variance when modeling LD tasks, it affords higher and more stable correlations with semantic similarity and relatedness norms, it captures context-dependent effects in the simulations performed on the MeMo dataset and on the stimuli from Till et al. (1988), and better preserves semantic relations in the replacement study. non-contextualised prototypes offer a particularly poor performance at deeper layers, where the influence of context on the resulting representation is highest. We hypothesize that this failure follows from the substantial difference between isolated words and the language on which BERT is trained, consisting of full passages. When the model cannot rely on any linguistic context, the relations it learned between lexical representations and context become useless, resulting in unreliable representations.

This observation poses a challenge to our claim that BERT provides a useful account of semantic memory, as the dynamic update of representations based on context fails when context is absent, unlike the behavior exhibited in psycholinguistic experiments for decades by participants presented with isolated words. Moreover, BERT shows much stronger empirical fit when considering exemplars and using them to compute new prototype representations that are not learned by the system, and which are not an inherent part of the model architecture but rather a post-hoc addition. Does this mean that the only way for BERT to provide adequate accounts of relevant psycholinguistic tasks depends on the storage of token-level representations (which are required to compute these type-level prototype representations) after all? We believe the answer is no: such token-level, context-dependent representations need not be stored, but they can be the result of simulations triggered when encountering a word. In line with what has been proposed in the embodied cognition literature (Barsalou, 1999; Dove, 2014; Fischer & Zwaan, 2008; Glenberg & Kaschak, 2002), conceptual understanding is mediated by simulations of the interactions with the concept (via perception, proprioception, motor interactions and more). We suggest that context-dependent lexical representations arise due to simulations which also involve the linguistic contexts in which a word is encountered[18]. Much like embodied simulations capitalize on interactions with a given object, we propose that conceptual understanding similarly leverages encounters with a word in linguistic contexts. The situational context in which a word has been experienced, even in the lack of explicit linguistic context in the here and now, will thus interact with the abstract semantic representation to yield different simulations of the word embedded in context, which will yield the context-dependent,

[18] We don't make any claim here with respect to the role of embodied simulations in conceptual understanding: By using BERT representations we focus purely on the linguistic side of conceptual understanding, which may exist next to embodied simulations.

token-level representations we used to derive average prototypes, in line with prepositional entropy effects on lexical decision latencies uncovered in Baayen et al. (2011). The fact that these representations work better than the type-level, abstract, context-independent prototype representations BERT learns suggests that this process of simulation which considers linguistic interactions with a word fosters richer, more informative representations and underlines the importance of learning a model of the interaction between invariance in experience and context rather than only the former. Why, then, does presenting isolated words yield worse results? And what does this tell us about BERT as a model of semantic memory, if we accept the analogy between isolated word tasks and showing a single word to the model?

It is useful to keep in mind that BERT's output layer relies on the deepest transformer encoder, hence it depends on the representation which results after a token has traversed all layers. The output the model provides at the deepest layer when presented with an isolated word, as we have shown, is often inappropriate and, especially when considering semantic similarity and relatedness, indicates a rather uninformative semantic space. However, it is unclear how the model could support an output based on intermediate layers, where sensible semantic representations have been observed across tasks also for non-contextualised embeddings. All these remarks call into question the straightforward and literal adoption of BERT as a fully-fledged model of semantic memory at the algorithmic level of description; rather, we argue that it should be seen as a model at the computational level of (Marr, 1982): It provides an account of *what* the target system is doing, but not necessarily an exact account of *how* it is doing it in a mechanistic, step-by-step way. The emphasis, we contend, is on the importance of integrating invariance in experience (resulting in useful type-level representations) with an abstract model of how variability in experience interacts with invariance (resulting in token-level representations).

BERT's objective function is rather narrow, aiming to optimize prediction of masked words and next sentences. This places a large emphasis on the role of a broad linguistic context, which differs from many language experiences humans have, where linguistic context may not be as extensive (but situational context is arguably far richer). We hypothesize that this strong focus on broad linguistic context leads the model to develop unreliable and unstable semantic representations at deeper layers in cases when context is absent. While prediction certainly plays a role in language comprehension (Willems et al., 2016) and acquisition (Fazekas et al., 2020; Vujović et al., 2021), the specific implementation BERT uses is limited and appears to be a prime difference between the model and the processes humans rely upon when deriving semantic representations. More broadly, prediction has been called into question as the single mechanism supporting language comprehension

(Huettig, 2015; Huettig & Mani, 2016). And even within the framework of predictive processing, prediction of upcoming lexical material may be just a part of the puzzle: recent work which relies on the Sentence Gestalt model suggests that a model trained to minimize prediction errors not only for upcoming lexical items but also for semantic roles offers a better account of the N400 (Lopopolo & Rabovsky, 2021), calling for more studies on the cognitive plausibility of training objectives and the introduction of new ones which may better reflect how humans learn.

We take these observations on the specific structure of BERT (different layers, unclear status of output based on intermediate representations) and its training objective (masked word prediction) as a cautionary tale against drawing strict equivalences between BERT layers and language processing steps, and against conceptualising BERT as an algorithmic account of semantic memory (Marr, 1982). On the contrary, we believe BERT - and the architecture it relies upon - can be a fruitful computational characterization of semantic memory, with a particular focus on how context-independent abstractions and context-dependent representations interact with each other through a model of how experience consists of invariant patterns embedded in an ever changing flux, which influences such invariant patterns and at the same time is constrained by them.

The emerging picture is thus that of a highly dynamic system, consisting of a set of static representations which are learned from exposure and constrain the malleability of the semantic space itself. Representations are flexible and can shift, but cannot be entirely superseded by the limited context used in our experiments (typically a single sentence). This lines up with behavioral evidence from Nieuwland and Van Berkum (2006) suggesting that many salient semantic features can be overwritten provided sufficient context, to the point that after having described a peanut with anthropomorphic traits, a sentence such as *The peanut was in love* elicits no semantic violation, whereas a sentence such as *The peanut was salted* does, in spite of the second being a prototypical feature of peanuts as we know them. Cevoli, Watkins, Gao, et al. (2022) provide exploratory evidence that when exposed to the same context participants read in Nieuwland and Van Berkum (2006)'s experiment, BERT also predicts verbs which would normally fit a person rather than a peanut, suggesting that the underlying semantic representation has shifted. In addition, a recent study by Michaelov, Coulson, et al. (2023) on the same item material shows that transformer-based language models such as BERT can well account for the effects observed by Nieuwland and Van Berkum (2006). Whereas Nieuwland and Van Berkum (2006) took their evidence as ruling out a stable semantic core, the success of BERT and similar language models in accounting for their data (Cevoli, Watkins, Gao, et al., 2022; Michaelov, Coulson, et al., 2023), as well as our own evidence (pointing to a different ease with which representations

can be morphed to fit a certain context), suggest that the lack of a semantic violation when reading of an amorous peanut can be explained by the tight interplay between context and lexical representations (of which context also consists of, evidently). We contend that people do have a stable representation for the word *peanut* and also for the other words in the sentences used in the experiment: the interaction between them and the knowledge of how context operates on lexical representations allow people to form a different representation of the target word, which evokes certain semantic features and downplays others.

Limitations

Despite the encouraging evidence we provide about BERT offering a useful model of semantic memory which allows the study of isolated word effects, context-dependent effects, and their interaction within a single framework, there are a few caveats to keep in mind. First of all, BERT is a purely linguistic model, in the sense that it learns from language patterns alone (which has received considerable criticism from embodiment research; see the seminal article by Glenberg and Robertson, 2000, and C. R. Jones et al., 2022, for a more recent version of the same argument also considering BERT). While this is certainly a limitation, it is again not inherent to the class of models of which it is an instance (see Günther et al., 2019), which have already started to be enriched with perceptual information. For example, the CLIP (Contrastive Language-Image Pretraining) model (Radford et al., 2021) uses the same architecture as BERT (Vaswani et al., 2017) but jointly learns linguistic and visual representations from images and corresponding captions (C. Jones et al., 2024; Pezzelle et al., 2021). A recent stream of studies have started comparing purely linguistic representations with multi-modal representations, observing a general advantage for the latter ones in a variety of settings relating to language processing, especially for concrete words (Pezzelle et al., 2021; Tikhonov et al., 2023).

Moreover, while we considered a variety of tasks relating to words in isolation and in context, other paradigms in psycholinguistic research could enrich our understanding of the extent to which contextualised embeddings learned by LLMs are a useful account of semantic representations. We investigated semantic violations in terms of replacements rather than controlled semantic violations, known to elicit distinctive ERP components such as the N400 (Dudschig et al., 2016; Hagoort et al., 2004; Kutas & Federmeier, 2011), which could offer a more stringent test. Our study suggests that BERT succeeds at integrating lexical representations which are not too different from the original word found in contexts, whereas it struggles with more dissimilar representations: The representation of a replacement shifts towards the neighborhood of the target, but never quite gets there. On the psychological level, this discrepancy could well result in processing costs or difficulties in integrating

certain concepts in the current sentence representation (see Brouwer et al., 2017, and references therein). In line with this intuition, we observed that replacements' surprisal is positively correlated with the dissimilarity between replacement and original target (see Appendix Word probability in context). This also influences the similarity between a contextualised embedding at the deepest layer and the static embedding for the same word type at layer 0: the less related the replacement is to the original word, the more dissimilar the contextualised representation is from the static one. These costs (high surprisal and more difficult contextualisation) for unrelated words tie in with evidence on N400 and P600 effects (Hagoort et al., 1993; Kutas & Hillyard, 1980). Initial explorations suggest Language Models relying on contextualised embeddings hold promise in accounting for human processing data on semantic violations (Michaelov, Bardolph, et al., 2023), with model-derived measures being better predictors for N400s than human predictability judgments (Michaelov et al., 2022). Michaelov, Bardolph, et al. (2023) compared measures of predictability and of semantic relatedness in accounting for processing data elicited by sentences which did or did not present a semantic violation known to elicit an N400 component, finding that predictability explains more variance than contextual similarity. However, the former was operationalized with fully static embeddings: our results show that contextualised embeddings capture more fine-grained semantic effects, especially in context, and should thus be preferred for these analyses.

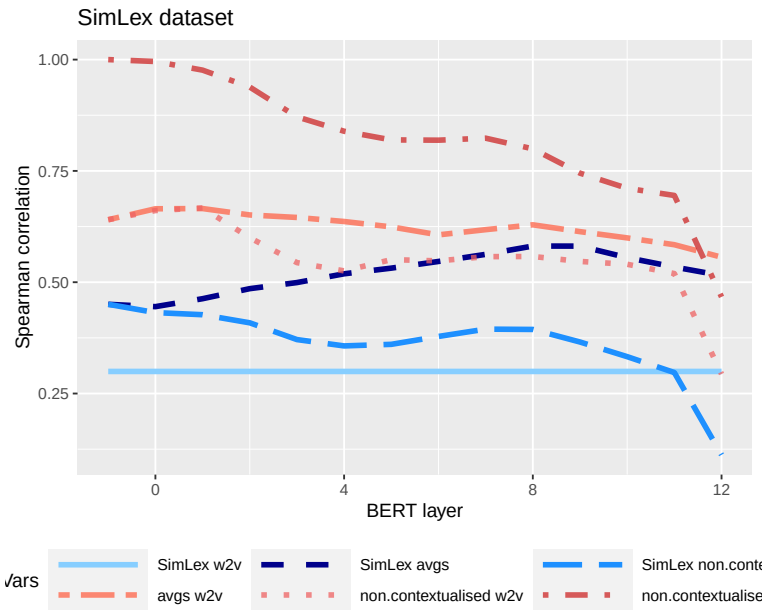
Conclusion

To sum up, we analysed BERT representations in a variety of settings, considering both contextualised, token-level and prototype, type-level representations in tasks which rely on isolated words and tasks which require the model to capture the interaction with context. We identified two main desiderata for psychologically valid semantic representations: i) that words have a semantic core which cannot be entirely overwritten by context and reflects how the word type is interpreted in isolation and ii) that context co-determines a word's representation when embedded in a sentence. Our analyses allow us to answer positively about BERT representations' validity and usefulness to study semantic effects in psycholinguistics. They do capture a stable semantic core which accounts for isolated word effects. Moreover, the word meaning can shift from this core following predictable and systematic context effects. We highlight that static and context-dependent representations co-exist in the same embedding space: in this, BERT offers a viable single implementation of both, and holds promise for analysing the interplay between prototypes and exemplars in a learning model (Buijtelaar & Pezzelle, 2023; Madabushi et al., 2022).

Acknowledgments

GC thanks Bruno Nicenboim for useful discussions on the scope of this paper. FG is supported by the German Research Foundation (DFG), Emmy-Noether grant "What's in a name?" (project number 459717703). FB is supported by the Hoffman–Yee Research Grants Program and the Stanford Institute for Human-Centered Artificial Intelligence. GA is partially supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA).

(a) *SimLex*



(b) *MEN*

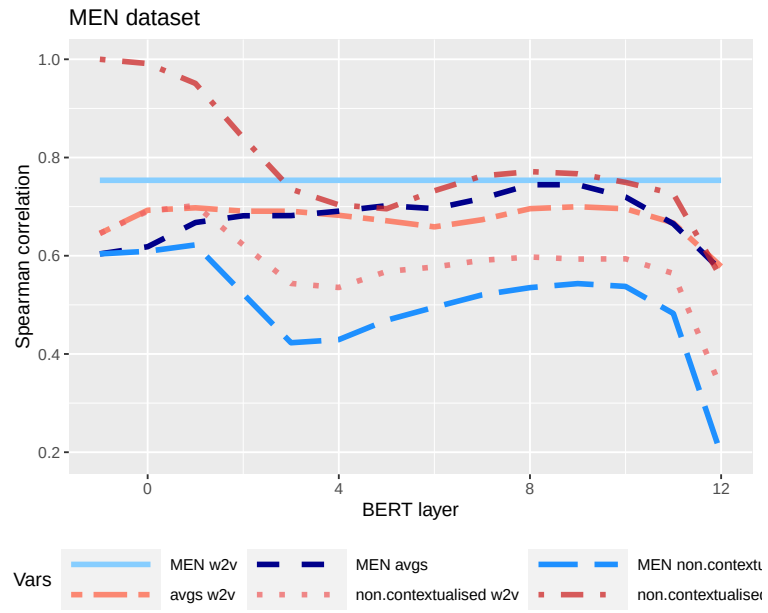
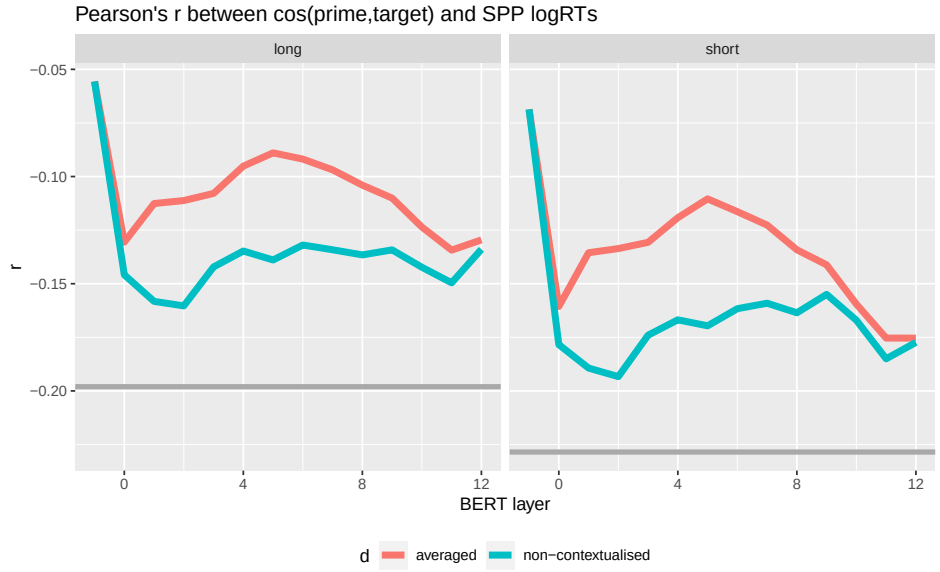


Figure 3

Correlations with semantic similarity and relatedness norms: SimLex (a) and MEN (b). We also plot the correlations between model-based cosine similarities, computed over the same word pairs. Correlations involving behavioral ratings are in shades of blue, correlations involving model-based cosine similarities are in shades of red.

(a) *Pearson's r*



(b) ΔAIC

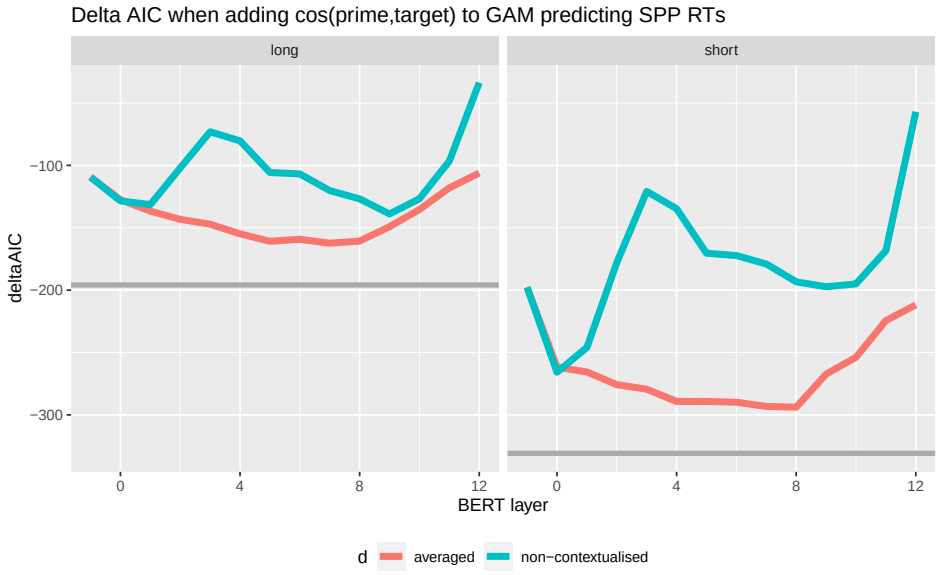


Figure 4

Prime-target cosine similarity computed from BERT embeddings and its effect on log Reaction Times (RTs) in the Semantic Priming Project (SPP) dataset. (a): Pearson's correlation between log RTs and BERT cos(prime, target); (b): ΔAIC of the model incorporating BERT cos(prime, target) with respect to the baseline multiple regression only including covariates. The x axis in each plot shows the BERT layers, and the color legend indicates the prototyping approach (red: averaged; cyan: non-contextualised). Gray lines indicate the score of the referenceword2vec model on each metric. Models were fitted separately for short and long Stimulus Onset Asynchrony (SOA), plotted as facets.

(a) *t statistic* (b) ΔAIC

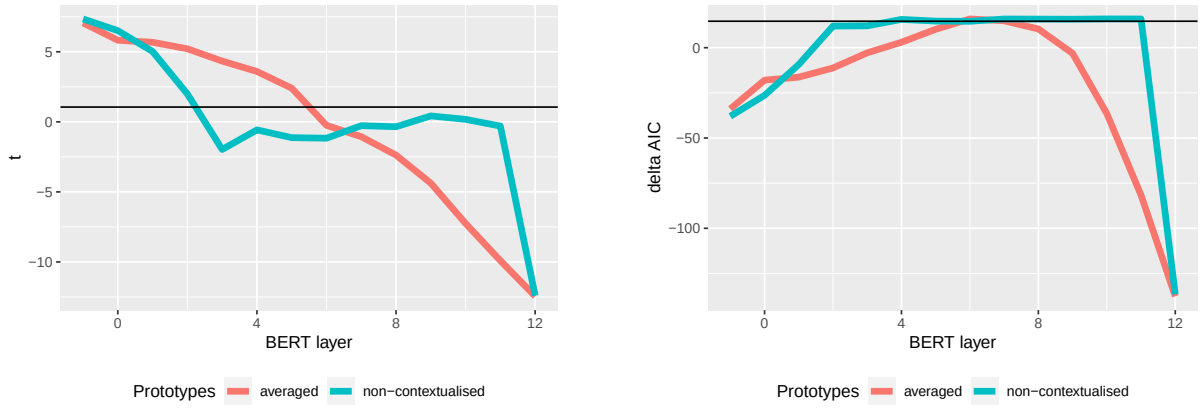


Figure 5

Semantic Neighborhood Density computed from BERT embeddings at different layers and its effect on Reaction Times (RTs) in the English Lexicon Project (ELP): (a): t statistic of SND as a predictor in a multiple regression model predicting RTs while controlling for other covariates; (b) ΔAIC of the model incorporating SND computed from BERT embeddings with respect to the baseline multiple regression only including covariates. The color legend indicates the prototyping approach (cyan: non-contextualised; red: averaged). Horizontal black lines indicate the score of the referenceword2vec model on each metric.

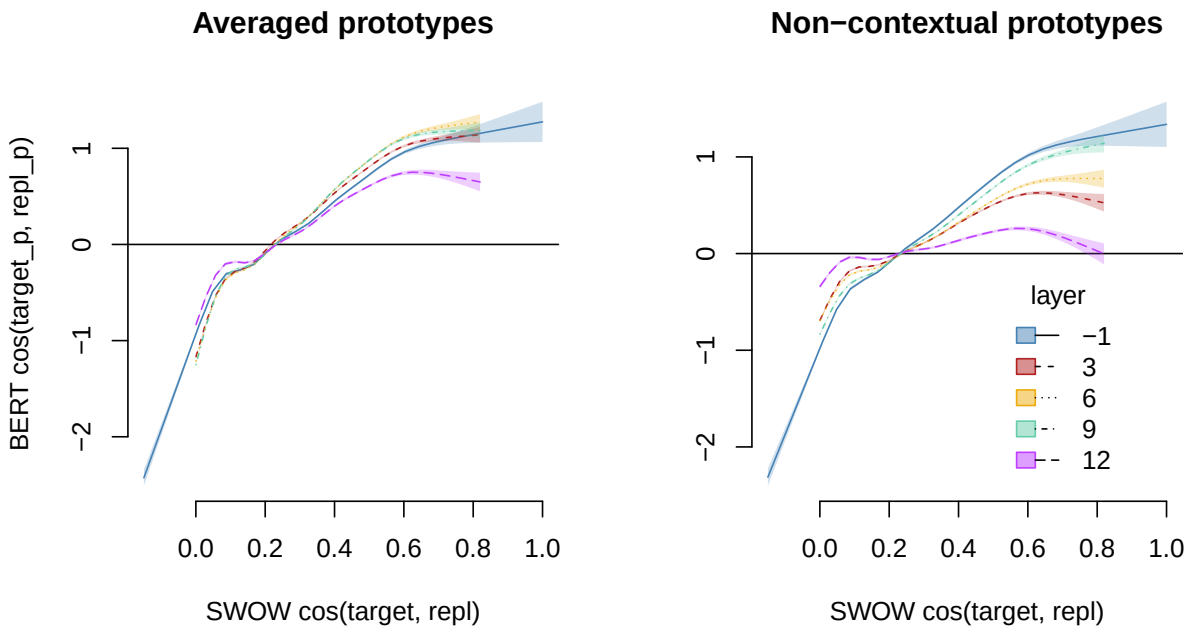


Figure 6

Predicted cosine similarity of prototype representations of target (target_p) and replacement (repl_p) in BERT (y-axis, fitted values) as a function of the similarity of the two words in a similarity space derived from the English Small World of Words (SWoW) associations using a random walk algorithm (x-axis, SWOW cos(target, repl)). Left: averaged prototypes; right: non-contextualised prototypes. Similarities in BERT are Box-Cox transformed and then z-standardised. The color and line legends identify BERT's layer. Both predictors are numerical and included in the model as simple smooths as well as in a partial tensor product.

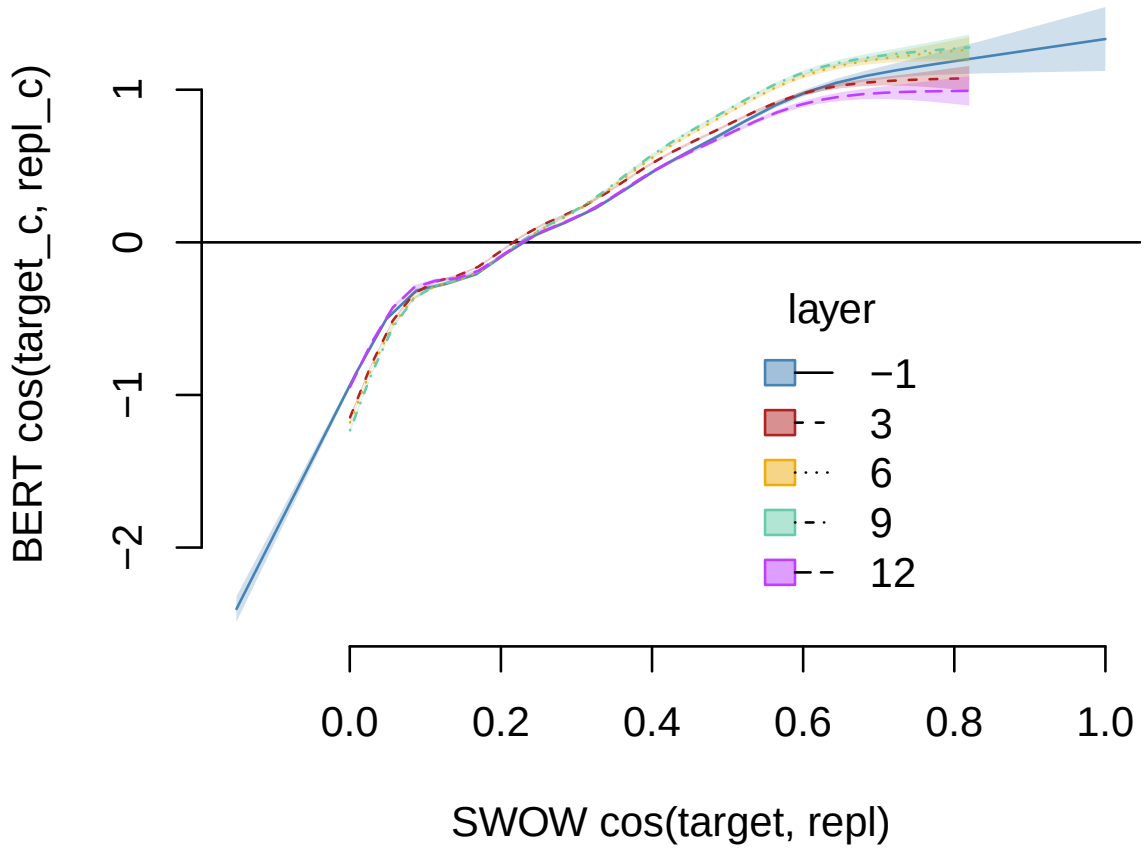


Figure 7

Predicted cosine similarity of contextualised representations of target (target_c) and replacement (repl_c) in BERT (y-axis, fitted values) as a function of the similarity of the two words in a similarity space derived from the English Small World of Words (SWoW) associations using a random walk algorithm (x-axis, SWOW cos(target, repl)). Similarities in BERT are box-cox transformed and then z-standardised. The color and line legends identify BERT's layer. Both predictors are numerical and included in the model as simple smooths as well as in a partial tensor product.

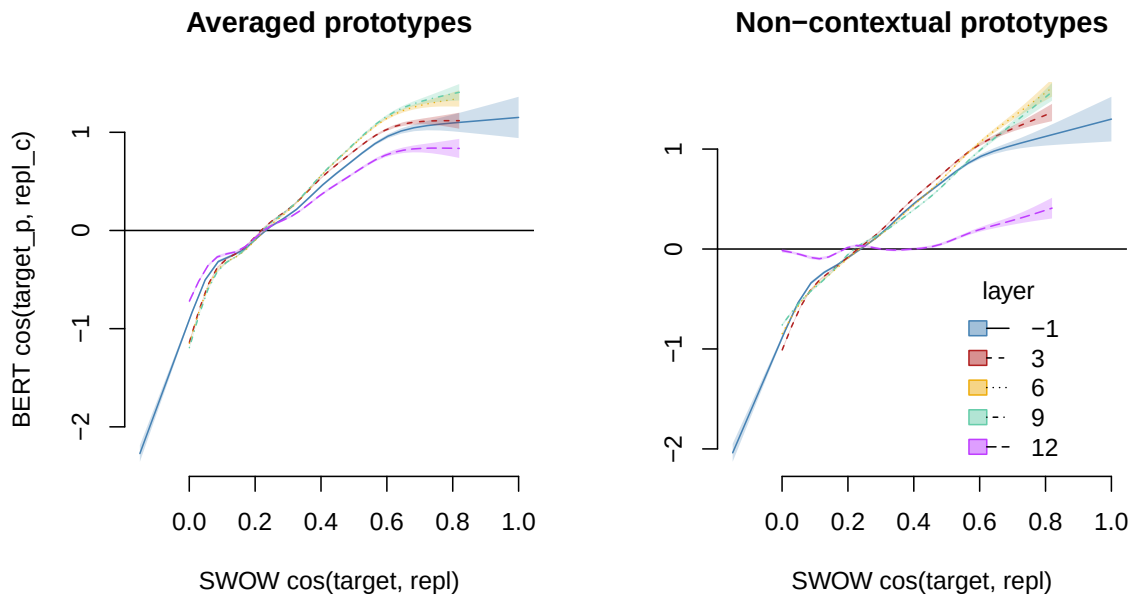


Figure 8

Predicted cosine similarity between target prototype (target_p) and contextualised replacement (repl_c). Left: averaged prototypes; right: non-contextualised prototypes. BERT cosine similarities (BERT cos$(\cdot,\cdot)$ have been box-cox-transformed and z-standardised. Lines show the predicted BERT cosine similarity between target and replacement as estimated by a GAM fitted with the similarity between target and replacement in the similarity space derived from the English Small World of Words (SWoW) free association data and BERT's layer as predictors (included as simple smooths and in a partial tensor product). Line and color legends identify the layers.

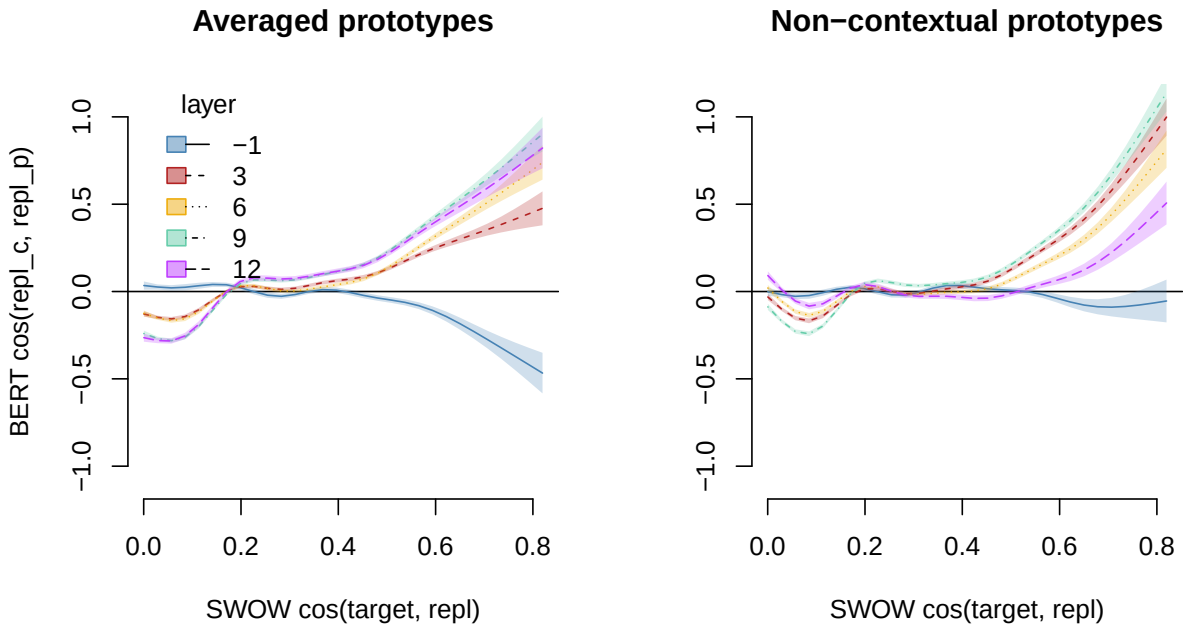


Figure 9

Predicted similarity in BERT (y-axis, BERT cos(·, ·), fitted values) between the prototype of the replacement (repl_p) and its contextualised representation (repl_c) as a function of the cosine similarity between target and replacement the similarity space derived from the English Small World of Words (SWoW) free association data (SWOW cos(target, repl)). The BERT cosine similarity was first Box-Cox transformed and then z-standardised. The color and line legends identify the layer. The underlying GAM models cosine similarity in BERT as a function of cosine similarity between target and replacement in the SWoW similarity space (simple smooth), layer (simple smooth), and a partial tensor product between the two.

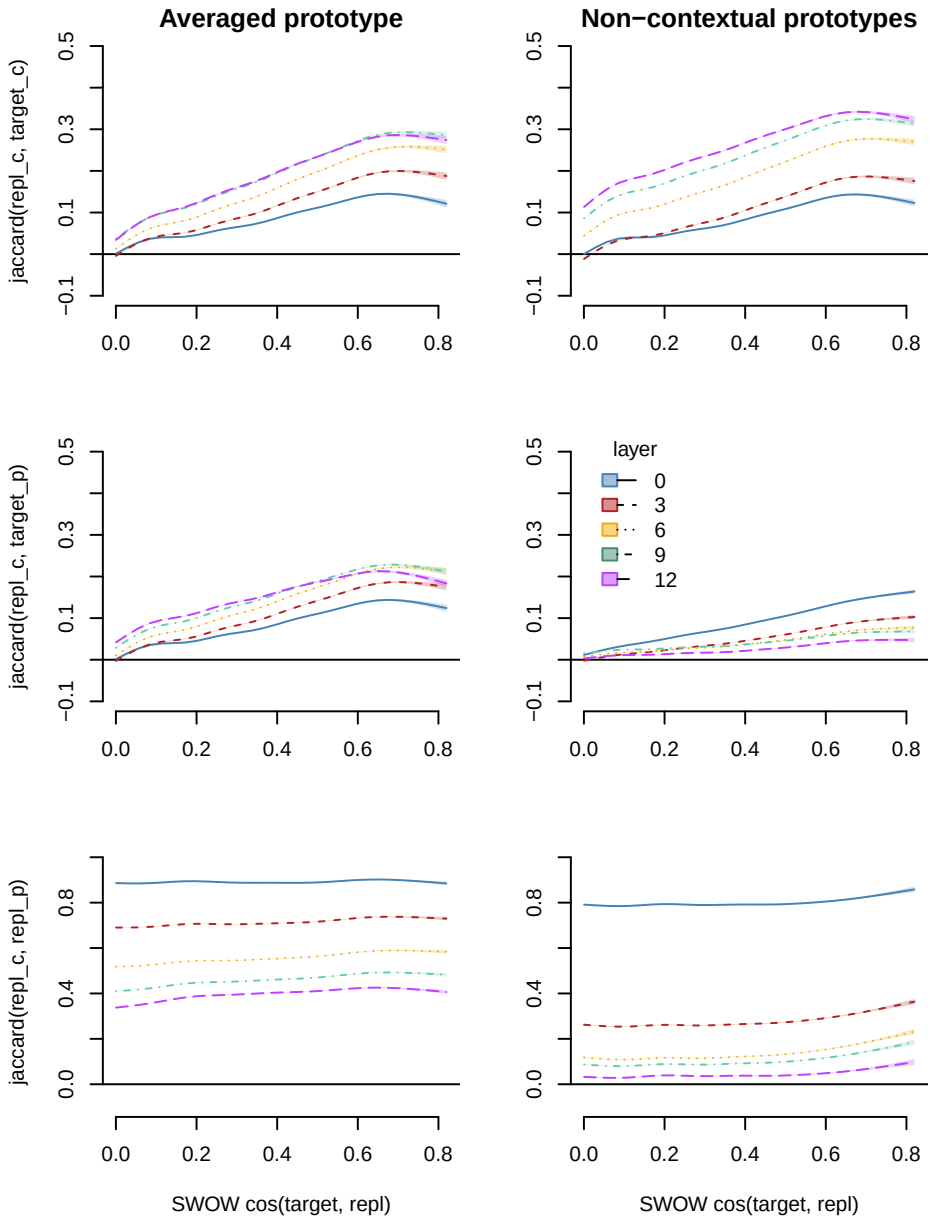


Figure 10

Jaccard coefficient between the nearest neighbors in BERT space of the contextualised replacement (repl_c) and the nearest neighbors of the the contextualised target (target_c, top), prototype target (target_p, mid), and prototype replacement (repl_p, bottom), shown separately for averaged (left) and non-contextualised (right) prototypes. The underlying GAMs model Jaccard coefficient (y-axis, fitted values) as a function of similarity between target and replacement inword2vec (simple smooth, x-axis), layer (simple smooth, color and line legends), and a partial tensor product between the two.

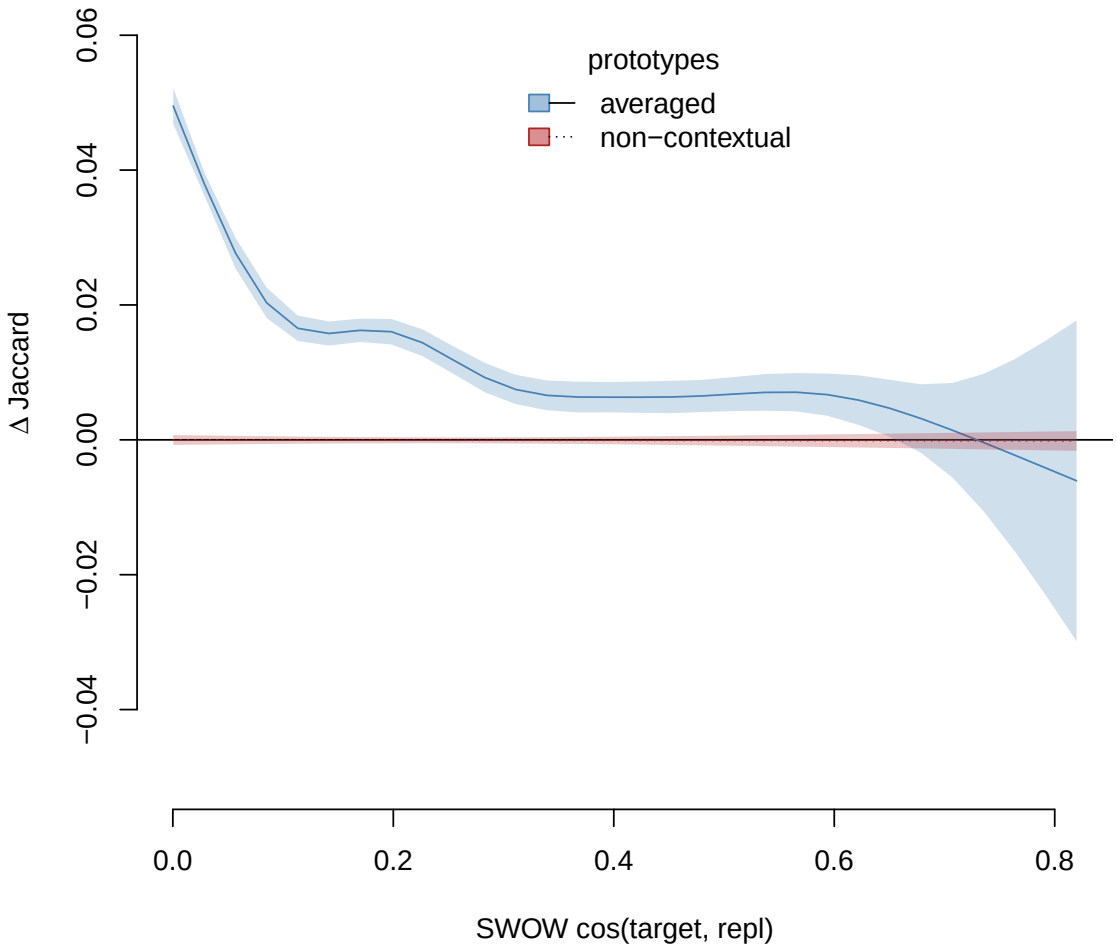
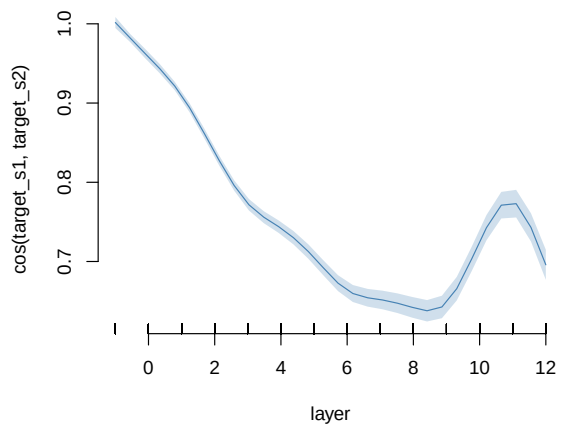


Figure 11

Predicted Δ (y-axis, fitted values) between (i) the Jaccard coefficient of the sets of nearest neighbors of the contextualised target (target_c) at layers -1 and 12, and (ii) the Jaccard coefficient between the sets of nearest neighbors of the contextualised replacement (repl_c) at layers -1 and 12, as a function of the cosine similarity in the SWoW similarity space (x-axis, SWOW cos(target, repl)). Color and line legends indicate the prototype method: blue, solid line for averaged prototypes; red, dotted line for non-contextualised prototypes. The underlying GAM models the Δ Jaccard coefficient $(jaccard(target_{c-1}, target_{c12}) - jaccard(replacement_{c-1}, replacement_{c12})$ *as a function of the cosine similarity between target and replacement in the SWoW similarity space (simple smooth). A predicted value of 0 indicates no difference, positive values indicate the target's neighbors are more coherent between superficial and deep layers.)*

(a) *same target, different sense*



(b) *similarity by condition*

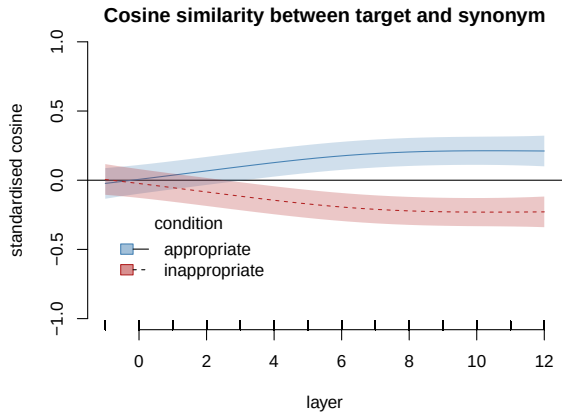
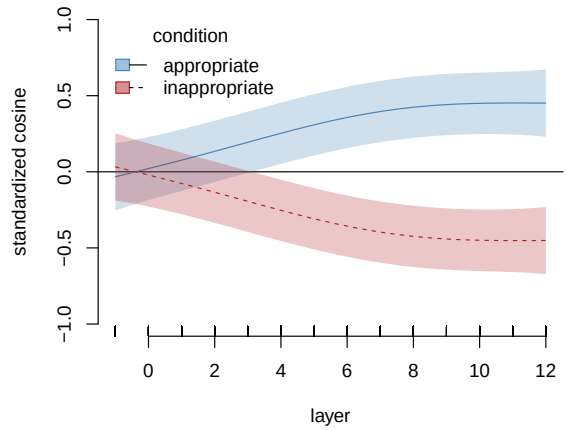


Figure 12

Word in Context (WiC) dataset. Effect of layer (x-axis) on (a) the cosine similarity (y-axis) between polysemous targets used in different sentences (s1 and s2) in different senses; and (b) the layer-wise standardized cosine similarity between polysemous target and appropriate synonym (blue, solid line) and between polysemous target and inappropriate synonym (red, dashed line), where the inappropriate synonym is the synonym of the target in the other sentence, and thus synonym of the target when used in a different sense.

(a) *averaging*



(b) *non-contextualised*

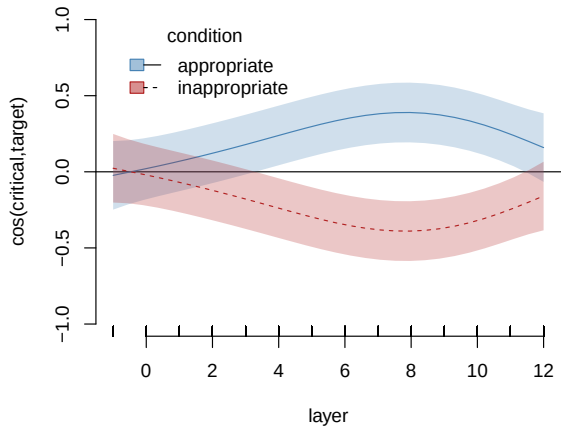


Figure 13

Priming in Context (Till et al., 1988). Effect of layer (x-axis) on the layer-wise z-standardized cosine similarity (y-axis, fitted values) between ambiguous critical word and target, plotted separately by target relation with the critical word (condition [appropriate - blue solid line | inappropriate - red dashed line]) for (a) averaged prototypes and (b) non-contextualised prototypes.

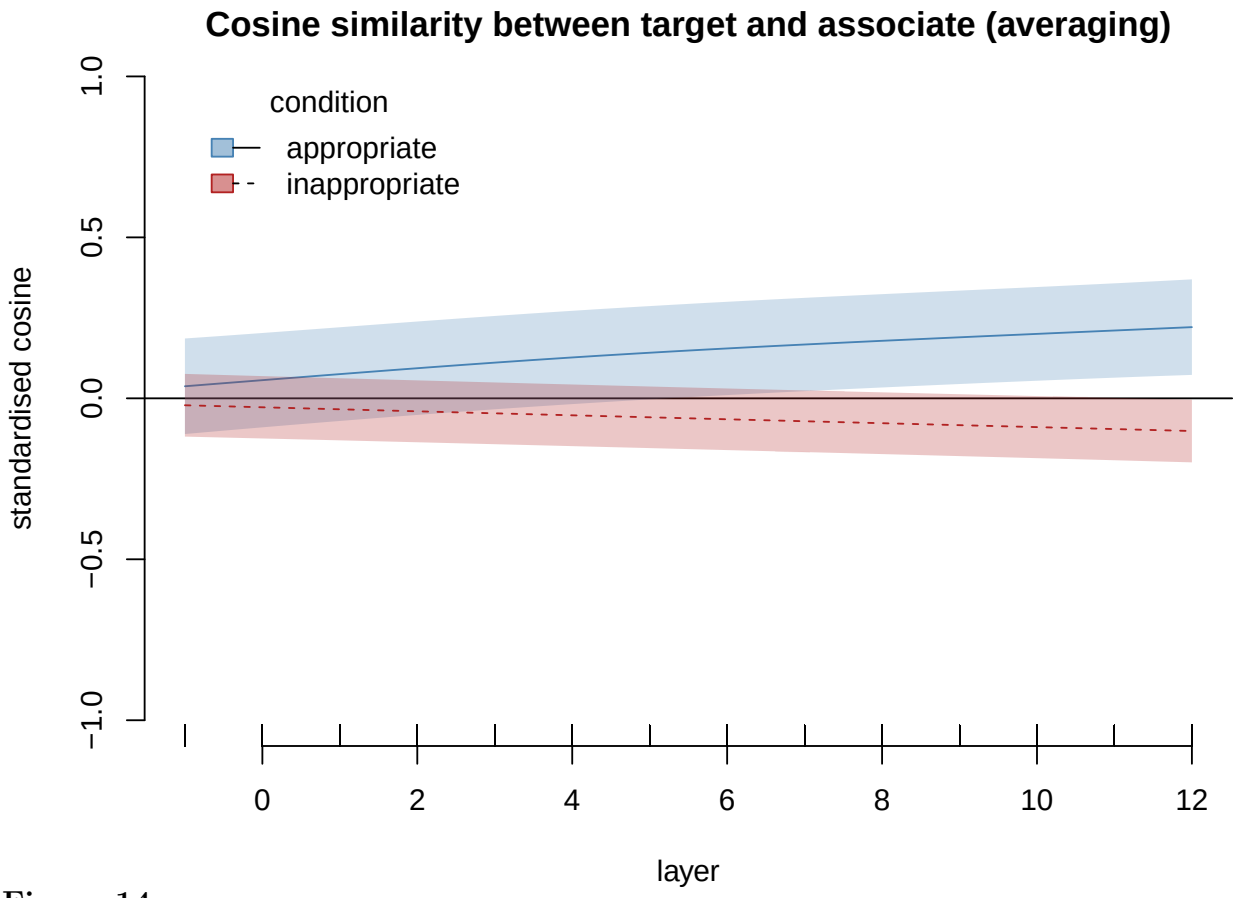


Figure 14

Meaning Modulation (MeMo) dataset. Layer-wise standardized cosine similarity (y-axis, fitted values) between the contextualised representation of the target word and the averaged prototype representation of the associate word over layers (x-axis), plotted separately for the two conditions: related (blue, solid line) indicates the target and associate word match; foil (red, dashed line) indicates the target and associate do not match.

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5-6), 509–559. https://doi.org/10.1177/0142723719869731

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 463–498.

Apidianaki, M. (2023). From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, *49*(2), 465–523. https://doi.org/10.1162/coli_a_00474

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 273–278).

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*, *31*(7), 940–966.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–481.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014a). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. https://doi.org/10.3115/v1/P14-1023

Baroni, M., Dinu, G., & Kruszewski, G. (2014b). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of ACL 2014*, 238–247.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 637–660.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*, 5418.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences*, *1505*, 55–78.

Beekhuizen, B., Armstrong, B. C., & Stevenson, S. (2021). Probing lexical ambiguity: Word vectors encode number and relatedness of senses. *Cognitive Science*, *45*(5), e12943.

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: A network science approach. *Thinking & Reasoning*, *23*, 158–183.

Bhatia, S., & Richie, R. (2022). Transformer networks of human conceptual knowledge. *Psychological Review*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

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

Bonandrini, R., Amenta, S., Sulpizio, S., Tettamanti, M., Mazzucchelli, A., & Marelli, M. (2023). Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, *145*, 101594.

Breit, A., Revenko, A., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2020). Wic-tsv: An evaluation benchmark for target sense verification of words in context. *arXiv preprint arXiv:2004.15016*.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A neurocomputational model of the n400 and the p600 in language processing. *Cognitive science*, *41*, 1318–1352.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, *49*(1), 1–47. https://doi.org/10.1613/jair.4135

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019a). Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, *51*(2), 467–479.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019b). Word prevalence norms for 62,000 english lemmas. *Behavior Research Methods*, *51*, 467–479.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, *44*(4), 991–997.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*, 1849–1863.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*, 531–544.

Buijtelaar, L., & Pezzelle, S. (2023). A psycholinguistic analysis of bert's representations of compounds. *arXiv preprint arXiv:2302.07232*.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in second language acquisition*, *24*(2), 215–221.

Bybee, J. L. (2013). Usage-based theory and exemplar representations of constructions.

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

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., & Jurafsky, D. (2022). Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, *119*(31), e2120510119.

Cassani, G., Bianchi, F., & Marelli, M. (2021). Words with consistent diachronic usage patterns are learned earlier: A computational analysis using temporally aligned word embeddings. *Cognitive Science*, *45*(4), e12963.

Cevoli, B., Watkins, C., Gao, Y., & Rastle, K. (2022). Shades of meaning: Natural language models offer insights and challenges to psychological understanding of lexical ambiguity.

Cevoli, B., Watkins, C., & Rastle, K. (2022). Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society Open Science*, *9*(6), 211837.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. *Attention and performance vi* (pp. 535–555). Routledge.

Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, *6*(3), 390–406.

Cramer, P. (1970). A study of homographs. In L. Postman & G. Keppel (Eds.), *Norms of word association* (pp. 361–382). Academic Press.

Daelemans, W., & Van Den Bosch, A. (2010). Memory-based learning. *The Handbook of Computational Linguistics and Natural Language Processing*, 154–179.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159–190.

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., & Navarro, D. (2016). Large-scale network representations of semantics in the mental lexicon. In M. N. Jones (Ed.), *Big Data in Cognitive Science: From Methods to Insights* (pp. 174–202). Routledge.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, *51*(3), 987–1006.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (cslb) concept property norms. *Behav Res Methods*, *46*(4), 1119–27. https://doi.org/10.3758/s13428-013-0420-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dove, G. (2014). Thinking in words: Language as an embodied medium of thought. *Topics in Cognitive Science*, *6*, 371–389. https://doi.org/10.1111/tops.12102

Dudschig, C., Maienborn, C., & Kaup, B. (2016). Is there a difference between stripy journeys and stripy ladybirds? The N400 response to semantic and world-knowledge violations during sentence processing. *Brain and Cognition*, *103*, 38–49.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint, arXiv:1909.00512*.

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? a registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*, *7*(11), 180877.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

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

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203.

Frisch, S. A. (2017). Exemplar theories in phonology. *The routledge handbook of phonological theory* (pp. 553–568). Routledge.

Gatti, D., Marelli, M., & Rinaldi, L. (2022). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, Advance online publication.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*, 558–565.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodies Theories of Meaning. *Journal of Memory and Language*, *43*, 379–401.

Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2021). Simple, interpretable and stable method for detecting words with usage change across corpora. *arXiv preprint arXiv:2112.14330*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–44. https://doi.org/10.1037/0033-295X.114.2.211

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

Günther, F., Dudschig, C., & Kaup, B. (2016a). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*, 626–653.

Günther, F., Dudschig, C., & Kaup, B. (2016b). Predicting lexical priming effects from distributional semantic similarities: A replication with extension. *Frontiers in Psychology*, *7*, 1646.

Günther, F., & Marelli, M. (2022). Patterns in caoss: Distributed representations predict variation in relational interpretations for familiar and novel compound words. *Cognitive Psychology*, *134*, 101471.

Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023). Vispa (vision spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*, *130*, 896–934.

Haber, J., & Poesio, M. (2024). Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, *50*(1), 351–417.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439–483.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438–441.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural shift or linguistic drift? comparing two computational measures of semantic change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*, *2016*, 2116–2121.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 75–86. https://doi.org/10.18653/v1/2020.cmcl-1.10

Harati, P., Westbury, C., & Kiaee, M. (2021). Evaluating the predication model of metaphor comprehension: Using word2vec to model best/worst quality judgments of 622 novel metaphors. *Behavior Research Methods*, *53*(5), 2214–2225.

Harris, Z. (1954). Distributional Structure. *Word*, *10*, 146–162.

He, P., Liu, X., Gao, J., & Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

Hendrix, P., & Sun, C. C. (2020). A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000819

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *NAACL-HLT*.

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153–172.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*, 665–695.

Hino, Y., Lupker, S. J., & Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 686–713.

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

Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory & Cognition*, *45*, 1350–1370.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research*, *1626*, 118–135.

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, *110*(2), 220.

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2023). A tutorial on open-source large language models for behavioral science. *PsyArXiv preprint*.

Hutchison, K. A., Balota, D. A., Cortese, M., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*, 1036–1066.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The semantic priming project. *Behav Res Methods*, *45*(4), 1099–114. https://doi.org/10.3758/s13428-012-0304-z

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*, 1099–1114.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*(2), 119–136. https://doi.org/10.1007/s42113-018-0008-2

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology*, *1*(3), 174–183.

Johns, B. T., Jamieson, R. K., Crump, M. J., Jones, M. N., & Mewhort, D. (2020). Production without rules: Using an instance memory model to exploit structure in natural language. *Journal of Memory and Language*, *115*, 104165.

Jones, C., Bergen, B., & Trott, S. (2024). Do Multimodal Large Language Models and Humans Ground Language Similarly? *Computational Linguistics*, 1–25. https://doi.org/10.1162/coli_a_00531

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distributional semantics still can't account for affordances. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. Busemeyer, Z. Wang, J. Townsend, & A. Eidels (Eds.), *Oxford Handbook of Mathematical and Computational Psychology* (pp. 232–254). Oxford University Press.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1470–1489.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, *95*, 163–182.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*, 257–266.

Kohl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of phonetics*, *21*(1-2), 125–139.

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition*, *9*, 587–598.

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

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*, 1065–1081.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, *44*(4), 978–990.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and cognition*, *65*(2), 145–168.

Lachmair, M., Dudschig, C., De Filippis, M., de la Vega, I., & Kaup, B. (2011). Root versus roof: Automatic activation of location information during word processing. *Psychonomic Bulletin & Review*, *18*, 1180–1188.

Lakens, D. (2022). Sample size justification. *Collabra: psychology*, *8*(1), 33267.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*, 195–208.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

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

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, *56*(4), 1269–1313.

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), e12955.

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Lopopolo, A., & Rabovsky, M. (2021). Predicting the n400 erp component using the sentence gestalt model trained on a large scale corpus. *bioRxiv*, 2021–05.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Lucas, M. (2000). Semantic priming without association. *Psychonomic Bulletin & Review*, *7*, 618–630.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 201–208.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 660–665). Erlbaum.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, *52*, 1271–1291.

Madabushi, H. T., Divjak, D., & Milin, P. (2022). Abstraction not memory: Bert and the english article system. *arXiv preprint arXiv:2206.04184*.

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

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of orthography-semantics consistency on word recognition. *The Quarterly Journal of Experimental Psychology*, *68*, 1571–1583.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*, 485–515.

Marr, D. (1982). *Vision: A computational approach*. Freeman & Co.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, 1–71.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements.

IEEE Transactions on Cognitive and Developmental Systems, *Advance online publication*. https://doi.org/10.1109/TCDS.2022.3176783

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2023). Can peanuts fall in love with distributional semantics?

Mickus, T., Paperno, D., Constant, M., & Van Deemter, K. (2020). What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. *Proceedings of the Society for Computation in Linguistics*, *3*, 235–245.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781v3*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

Milin, P., Tucker, B. V., & Divjak, D. (2020). A learning perspective on the emergence of abstractions: The curious case of phonemes. *arXiv preprint arXiv:2012.07499*.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408.

Murphy, G. L. (2002a). *The Big Book of Concepts*. MIT Press.

Murphy, G. L. (2002b). Conceptual Combination. In G. L. Murphy (Ed.), *The Big Book of Concepts* (pp. 443–475). MIT Press.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336). Erlbaum.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 859–864.

Nick Reid, J., & Katz, A. N. (2018). Vector space applications in metaphor comprehension. *Metaphor and Symbol*, *33*(4), 280–294.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *J Cogn Neurosci*, *18*(7), 1098–111. https://doi.org/10.1162/jocn.2006.18.7.1098

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*, 175–190.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. A. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt 2018)* (pp. 2227–2237). Association for Computational Linguistics.

Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word comprehension. *Journal of Memory and Language*, *117*, 104194.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*, 161–167.

Pezzelle, S., Takmaz, E., & Fernández, R. (2021). Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, *9*, 1563–1579.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Pilehvar, M. T., & Camacho-Collados, J. (2018). Wic: The word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, *4*(1), 86.

Rodd, J. M. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science*, *15*(2), 411–427. https://doi.org/10.1177/1745691619885860

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

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive science*, *28*(1), 89–104.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–46). Lawrence Erlbaum Associates.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Roy, A., & Pan, S. (2021). Incorporating medical knowledge in BERT for clinical relation extraction. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5357–5366. https://doi.org/10.18653/v1/2021.emnlp-main.435

Sahlgren, M. (2008). The Distributional Hypothesis. *Italian Journal of Linguistics*, *20*, 33–53.

Shah, R., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., & Yang, D. (2022). When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2322–2335. https://aclanthology.org/2022.emnlp-main.148

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using hidex. *Behavior Research Methods*, *42*, 393–413.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Harvard University Press.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). Lexops: An r package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*(6), 2372–2382.

Tikhonov, A., Bylinina, L., & Paperno, D. (2023). Leverage points in modality shifts: Comparing language-only and multimodal word representations. *arXiv preprint arXiv:2306.02348*.

Till, R. E., Mross, E. F., & Kintsch, W. (1988). Time course of priming for associate and inference words in a discourse context. *Memory & cognition*, *16*(4), 283–298.

Trott, S., & Bergen, B. (2023). Word meaning is both categorical and continuous. *Psychological Review*.

Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive science*, *35*(2), 251–296.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, *67*(6), 1176–1190.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2022). itsadug: Interpreting time series and autocorrelated data using gamms [R package version 2.4.1].

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*, 1176–1190.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. *Journal of Memory and Language*, *119*, 104231.

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., & Korhonen, A. (2020). Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, *46*(4), 847–897. https://doi.org/10.1162/coli_a_00391

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. https://doi.org/10.18653/v1/W18-5446

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45*(4), 1191–1207.

Wilcox, E., Vani, P., & Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 939–952. https://doi.org/10.18653/v1/2021.acl-long.76

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*, 742–750.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic bulletin & review*, *15*(5), 971–979.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic bulletin & review*, *23*(4), 1015–1027.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*, 168–171.

Appendix A
Word probability in context

In this supplementary analysis, we study the relation between the semantic relation between target and replacement in the Small World of Words similarity space on the one hand, and word probability in context on the other. While our work focuses on an analysis of representations, it is useful to double check that these representations influence predictions of masked word in a sensible way.

We thus target the log probability of the replacement in context (Ettinger, 2020), which gives a first indication of how sensible the underlying semantic space is. We first averaged the probability for each target-replacement pair over the 50 sentences per pair. We then fitted a GAM predicting the log probability as a function of the cosine similarity in the SWoW similarity space, the frequency of the target, and the frequency of the replacement, all entered as simple smooths). Frequency estimates were taken from SUBTLEX-US (Brysbaert et al., 2012, accessed via LexOPS Taylor et al., 2020) and were included as we expect the probability to also depend on how frequent the target and replacement are. A very frequent target should yield lower probabilities for any replacement, while a frequent replacement should yield a higher probability no matter what target it replaces. We went a step further and included partial tensor products (i.e., interactions) between cosine similarity and target frequency, and between cosine similarity and replacement frequency to investigate whether the frequency of target and replacement modulates the effect of cosine similarity. Table A1 summarizes the coefficients, while Figure A1 shows the effects visually.

Predictor	edf	Ref.df	F	p
s(cos sim)	7.771	8.614	2453.439	<.001
s(target f)	6.201	7.295	428.370	<.001
s(replacement f)	5.671	6.869	514.145	<.001
ti(cos sim, target f)	10.571	12.411	13.470	<.001
ti(cos sim, replacement f)	7.774	9.479	9.736	<.001

Table A1

Coefficients of the GAM fitted to predict the probability of the replacements in context as a function of cosine similarity (cos sim) between target - the word which was replaced - and replacement - the word used to replace it, the frequency of the target (target f), the frequency of the replacement (replacement f), and partial tensor products between cosine similarity and target frequency, and between cosine similarity and replacement frequency. The table provides the estimated degrees of freedom (edf), the reference degrees of freedom (Ref.df), the F statistic and the corresponding p-value.

Our predictions are borne out: the probability of a replacement increases when the cosine similarity in the SWoW similarity space between replacement and target increases; it

also increases when the frequency of the replacement increases but drops when the frequency of the target increases. The interaction between target frequency and cosine similarity shows that, at low frequencies, the effect is entirely reducible to cosine, with low probabilities being predicted when the cosine similarity is particularly low. At higher target frequencies, however, we see that predicted probabilities are low even at non-negligible similarity values, suggesting that when a target is very frequent, the replacement is predicted at low probabilities even if it is semantically more similar, as expected. Turning to the interaction between replacement frequency and semantic similarity we see a similar, albeit weaker, pattern in the opposite direction: when the replacement is frequent, the predicted probability is higher even at more moderate values of similarity, again as per expectations. This analysis confirms that BERT embeddings respond in the expected way to the interplay between frequency and semantic fit.

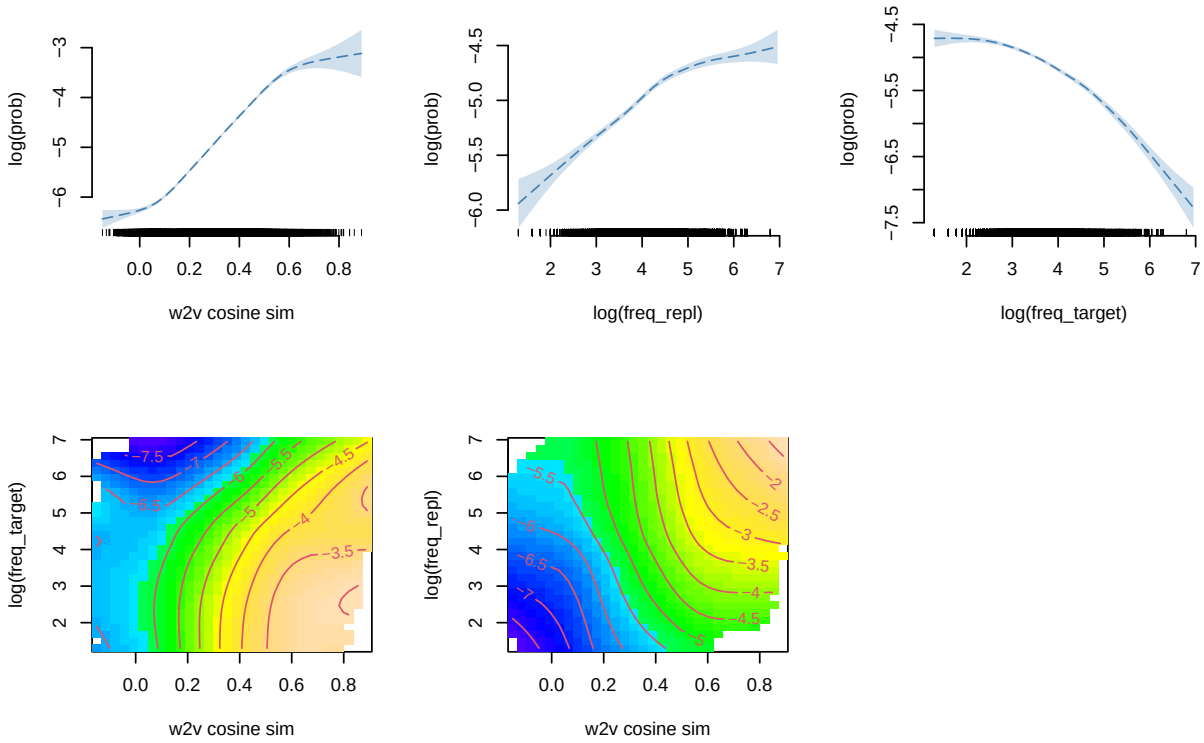


Figure A1

Effects of word2vec similarity, replacement frequency, target frequency and interactions between layer and frequency measures on log probability of the replacement in English sentences originally containing the target (y-axis) as estimated in a GAM where each predictor was included as a simple smooth and interactions were included as non-linear partial tensor products. The top-left panel shows the effect of cosine similarity in the reference word2vec space (x-axis) on log probability (y-axis, fitted values excluding random effects); the top-mid panel shows the effect of replacement frequency (x-axis) on log probability (y-axis, fitted values excluding random effects); the top-right panel shows the effect of target frequency (x-axis) on log probability (y-axis, fitted values excluding random effects); the bottom-left panel shows the partial tensor product between cosine similarity (x-axis) and target frequency (y-axis) on log probability (color); the bottom-mid panel shows the partial tensor product between cosine similarity (x-axis) and replacement frequency (y-axis) on log probability (color).

Appendix B
MeMo dataset reliability analysis

In order to assess how robust the interaction between condition [appropriate | inappropriate] and layer is when analysing meaning modulations within a same sense, we adopted a bootstrapping approach. We sampled 20, 30, and 40 target words without replacement, and subset the MeMo dataset to retrieve only the observations pertaining to the target words in the sample, including both appropriate and inappropriate associates. We then fit a Linear Mixed Effect Regression (LMER) model with the cosine between the contextualised embedding of the target word embedded in the corresponding sentence and the prototype embedding of the associate. We fit two separate models for the two prototyping methods, averaged and non-contextualised. From this model, we derived the beta coefficient of the interaction term and the corresponding t statistic. This analysis allows us to directly check the direction of the effect as well as the corresponding statistic. Finally, we fit two GAMMs predicting the same dependent variable as the LMER model. The first GAMM includes the following independent variables: condition as a parametric term and layer as a simple smooth. The second GAMM further includes the interaction between the two terms. We then used the *compareML* function from the *itsadug* package to test whether the second model has a better fit than the first, i.e., whether the interaction improves model fit quantified as the ΔAIC. We derived 100 samples, and performed the same analyses on each. Finally, we visualize the distribution of the three statistics of interest (beta coefficient of the target interaction and corresponding t statistic from the LMER, and the ΔAIC from the GAMMs). Figure B1, Figure B2, and Figure B3 show the distribution of the beta coefficient, t statistic, and ΔAIC respectively.

These analyses show that with the current sample size of target words and associates we are unlikely to overestimate the effect reported for averaged prototypes. When considering 20 target words we do observe larger variability, showing that certain target words emphasize the effect while others hinder it, but even the sample yielding the smallest beta coefficient ($\beta = 0.0136$) yields a robust effect in the predicted direction ($t = 3.666$) and including the interaction in the GAMM improves model fit ($\Delta AIC = 12.42$). In the same vein, the sample yielding the smallest t statistic ($t = 3.646$) also yields an interaction coefficient in the predicted direction ($\beta = 0.0158$) and shows an improvement in model fit after including the target interaction ($\Delta AIC = 12.26$, which is also the smallest increase in model fit). These are however the worst case scenarios, suggesting that the reported effects are robust and that the chosen sample size ensures a decent power. We hope this dataset sparks interest in such subtle meaning modulations and we look forward to attempts to replicate the effect with different target words, associates, and sentences.

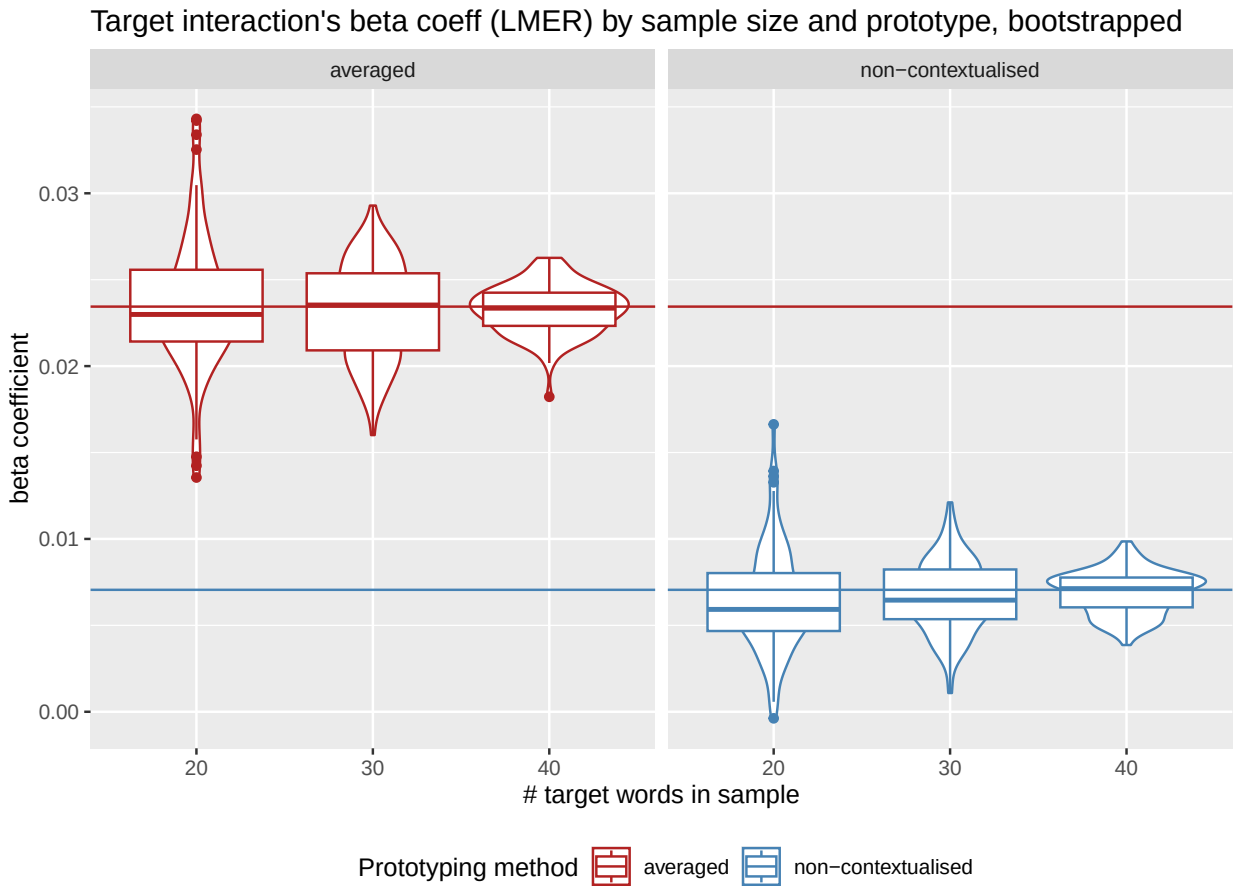


Figure B1

*Meaning Modulation (MeMo) dataset. Distribution of the beta coefficient of the target interaction, condition*layer, estimated fitting 100 Linear Mixed Effects Regression (LMER) models over 100 random samples of 20, 30, and 40 target words. y axis: value of the beta coefficient; x axis: number of target words in each sample; color and faceting: prototyping method, averaging or non-contextualised. Horizontal lines indicate the value of the beta coefficient estimated using the full MeMo dataset.*

At the same time, we observe that when using non-contextualised prototypes the effect is unlikely to be there and also that with a smaller sample size and a certain subset of target words we may have incorrectly concluded that the target interaction influenced the similarity patterns also with this prototyping method. The highest beta coefficient ($\beta = 0.0166$) has a corresponding t statistic of 3.054, which would have been significant under a LMER analysis. The full MeMo dataset, however, provides sufficient robustness to rule out that the target interaction has a reliable effect on non-contextualised prototypes.

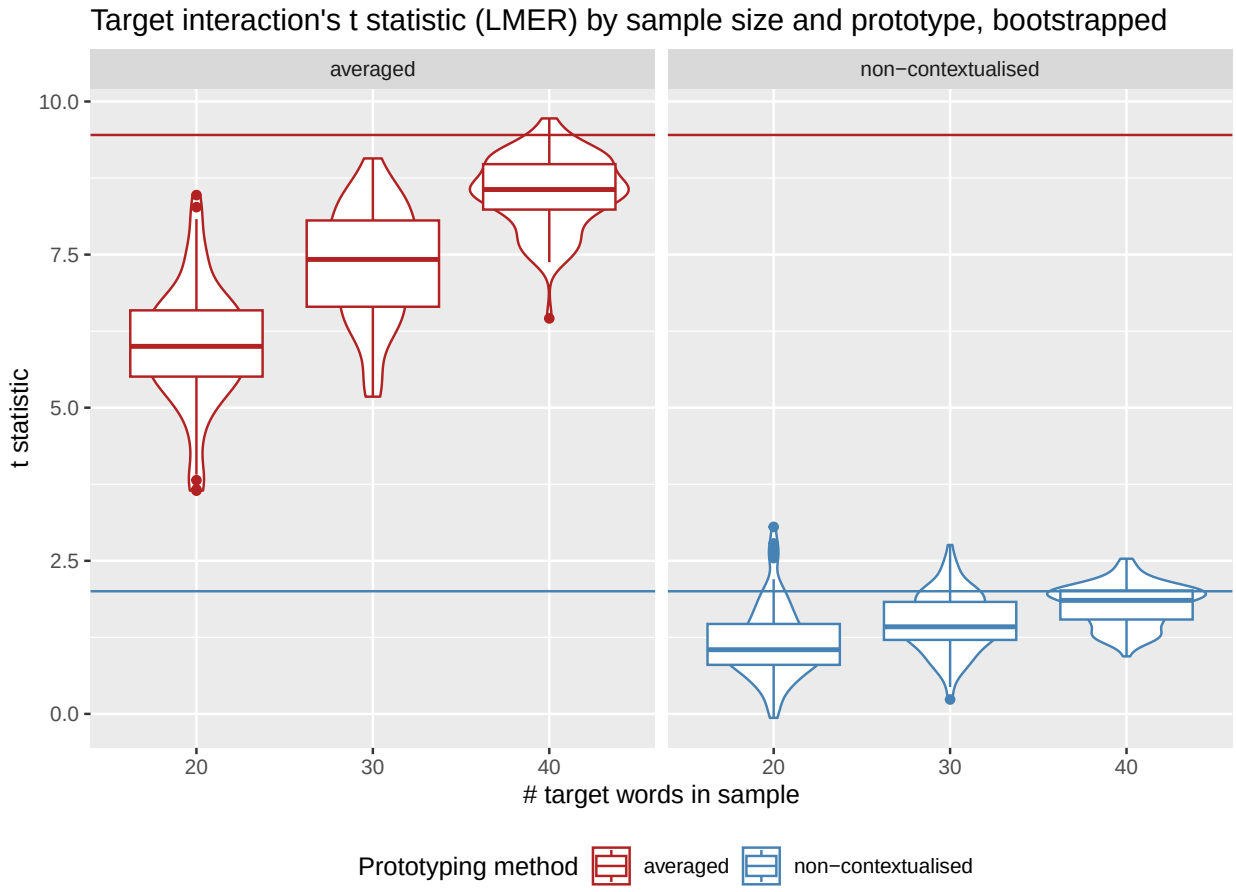


Figure B2

*Meaning Modulation (MeMo) dataset. Distribution of the t statistic of the target interaction, condition*layer, estimated fitting 100 Linear Mixed Effects Regression (LMER) models over 100 random samples of 20, 30, and 40 target words. y axis: value of the t statistic; x axis: number of target words in each sample; color and faceting: prototyping method, averaging or non-contextualised. Horizontal lines indicate the value of the t statistic estimated using the full MeMo dataset.*

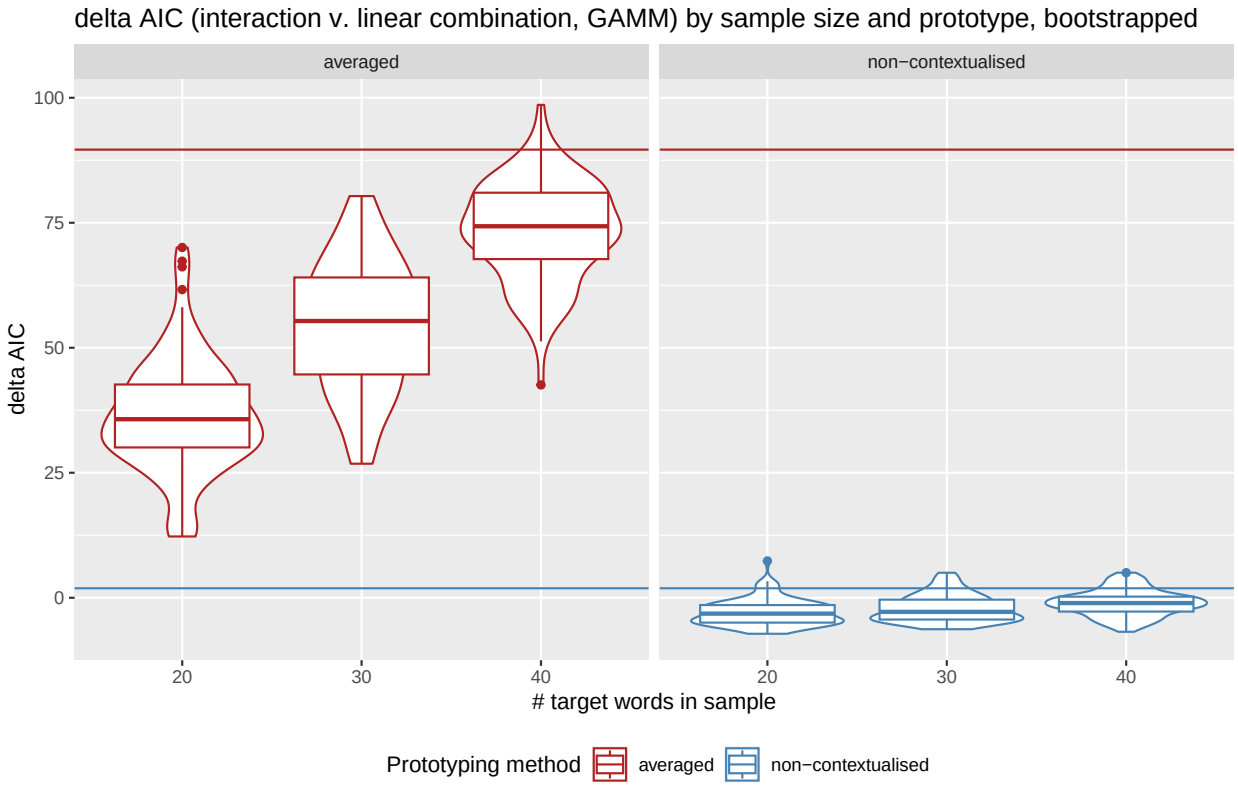


Figure B3

Meaning Modulation (MeMo) dataset. Distribution of the ΔAIC between a GAMM including and a GAMM not including the target interaction, layer by condition, over 100 random samples of 20, 30, and 40 target words. y axis: value of the ΔAIC; x axis: number of target words in each sample; color and faceting: prototyping method, averaging or non-contextualised. Horizontal lines indicate the value of the ΔAIC estimated using the full MeMo dataset.

Appendix C
Mathematical details of the Instance Theory of Semantics (ITS model

- Each letter string (word or nonword) is an n-dimensional vector w. Each dimensional value is sampled from $N(0, \frac{1}{n})$. This vector is assumed to just represent the form of the word, and does not yet carry any specific semantic information (the expected similarity of these vectors for any word pair is 0)

- The memory of a given language example (like a word, sentence, or document), called an *instance context* c_i, is the vector sum of all h words in that example:

$$c_i = \sum_{j=1}^{j=h} w_{ij} \tag{C1}$$

- All memories for each individual instances are stored in a matrix M, where each row is the *memory trace* of one language example (so $M_i = c_i$). Note that this matrix M can have a very large number of rows, because it has a memory trace of each language example that the individual has ever encountered. The number of columns is always n

- All information is stored in M. There are no stored representations of word meanings, only memories of individual language examples. Word meanings are created during *retrieval*, where they are referred to as *echo*

- During retrieval, a word (that would be encountered by the speaker) serves as a *probe* p into memory. This probe generates an *echo* from memory, in a two-step process:

    1. The probe activates *all* traces in memory M. The amount of activation of each trace is a function of the cosine similarity between probe and trace (more specifically, the cube):

$$a_i = cos(p, M_i)^3 \tag{C2}$$

    (as the cube of a cosine, a_i is bound between -1 and 1). By computing the cube, only *very similar* traces will receive relevant amounts of activation. This is relevant since there can be a very large number of traces, most of which will be completely irrelevant to the probe

2. Then, *all activated traces* are summed up via weighted addition to produce the *echo*. The weights are the activation values. So for each dimensional value j in echo e we have:

$$e_j = \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} a_i \times M_{ij} \tag{C3}$$

, with m being the number of memory traces in M, and M_{ij} being dimensional value j of memory trace M_i

3. The echo is the semantic representation retrieved for the probe

- The similarity between two echoes is simply calculated as their cosine similarity $cos(e_1, e_2)$
- It is also possible to present a *joint probe* consisting of more than one word. In that case, the general retrieval process still works as presented above. The only difference is the computation of activation: In that case, the formula in (C2) is generalized to

$$a_i = \prod_{k=1}^{k=h} cos(p_k, M_i)^3 \tag{C4}$$

, with p_k being the kth word in the probe (note that for $k = 1$, (C4) is equivalent to (C2)). This will, for example, produce a different echo for the probe [*break*, *glass*] than for [*break*, *car*]