Compositionality and Sentence Meaning: Comparing Semantic Parsing and Transformers on a Challenging Sentence Similarity Dataset

James Fodor[*]
The Centre for Brain, Mind and Markets
The University of Melbourne

Simon De Deyne[**]
School of Psychological Sciences
The University of Melbourne

Shinsuke Suzuki[†]
Faculty of Social Data Science
Hitotsubashi University

One of the major outstanding questions in computational semantics is how humans integrate the meaning of individual words into a sentence in a way that enables understanding of complex and novel combinations of words, a phenomenon known as compositionality. Many approaches to modelling the process of compositionality can be classified as either 'vector-based' models, in which the meaning of a sentence is represented as a vector of numbers, or 'syntax-based' models, in which the meaning of a sentence is represented as a structured tree of labelled components. A major barrier in assessing and comparing these contrasting approaches is the lack of large, relevant datasets for model comparison. This paper aims to address this gap by introducing a new dataset, STS3k, which consists of 2,800 pairs of sentences rated for semantic similarity by human participants. The sentence pairs have been selected to systematically vary different combinations of words, providing a rigorous test and enabling a clearer picture of the comparative strengths and weaknesses of vector-based and syntax-based methods. Our results show that when tested on the new STS3k dataset, state-of-the-art transformers poorly capture the pattern of human semantic similarity judgments, while even simple methods for combining syntax- and vector-based components into a novel hybrid model yield substantial improvements. We further show that this improvement is due to the ability of the hybrid model to replicate human sensitivity to specific changes in sentence structure. Our findings provide evidence for the value of integrating multiple methods to better reflect the way in which humans mentally represent compositional meaning.

∗ Email: fods12@gmail.com

∗∗ Email: simon.dedeyne@unimelb.edu.au

† Email: shinsuke.szk@gmail.com

1. Introduction

An important goal of computational semantics is to develop formal models to describe how humans understand and represent the meaning of words and sentences (Hampton 2017; Boleda 2020). There are many aspects of meaning these theories seek to capture, including the descriptive content of words (dictionary definitions), the relationship between language and the external world (truth conditions), fuzzy gradations of meaning (polysemy), and hierarchical relations between words (hyponymy)(Boleda and Herbelot 2016; Emerson 2020). This paper focuses on how individual words are combined into sentences to express complex and potentially novel ideas, a phenomenon often referred to as *compositionality*. Specifically, we assess how well different classes of language models capture human compositional representation of sentence meaning, providing a scaffold for developing formal models of human compositionality.

Beginning with the work of Gottlob Frege (Frege et al. 1892), most accounts in theoretical linguistics have appealed to the principle of compositionality as essential to human processing of sentence meaning (Montague 1970; Fodor and McLaughlin 1990; Baroni, Bernardi, and Zamparelli 2014). Though the term has no single accepted definition, broadly speaking, a symbolic system is said to exhibit compositionality if the truth value of a composite expression is a function only of the symbols contained in that expression and the formal syntactic rules used to combine them (Pelletier 2017). It has been argued that compositionality explains the *productivity* of language, the ability to use rules and concepts to produce and understand sentences never previously encountered (Szabó 2020; Löhr 2017). For instance, we can understand "man bites dog" by understanding the relation between the subject, verb, and object, even if we have never heard of a man biting a dog before (Frankland and Greene 2020). Compositionality also explains the *systematicity* of language, whereby understanding a sentence entails the ability to understand systematic variants of that sentence (Amigó et al. 2022).[1] Given that humans are capable of understanding a vast array of rich, complex sentences which they have never before encountered, many commentators have argued that any adequate theory of semantics must be able to account for compositional generalisation (Fodor and Pylyshyn 1988; Baroni, Bernardi, and Zamparelli 2014; Boleda and Herbelot 2016; Frankland and Greene 2020).

Such considerations have motivated the development of several distinct approaches to representing sentence meaning. *Vector-based* semantics[2] derives from the distributional semantics tradition in which a word, sentence, or passage is represented as a vector of numbers, the direction of which in semantic space represents the meaning of that word or passage (Erk 2012; Clark 2015; Boleda 2020). Early approaches in this tradition were based on explicitly modelling the distribution of word occurrences in a corpus and using this to construct an embedding (Deerwester et al. 1990). More recent approaches instead train neural networks on tasks such as next word prediction (Mikolov et al. 2013), and hence are sometimes called neural network (Baroni 2020) or deep-learning representations (Pavlick 2022). Currently the most capable vector-

---

1 Hence, for instance, if we understand "John loves Mary", we necessarily understand "Mary loves John" (Baroni 2020), even though neither claim necessarily entails the other. In a compositional system, predicates and their arguments are represented independently, thereby allowing novel systematic variations of such arguments, (such as interchanging 'Mary' with 'John') to be understood (Martin and Baggio 2020).

2 See subsubsection 8.1.1 in the Appendix for further discussion of our terminology of *vector-based* and *syntax-based*.

based models are based on the transformer neural network architecture, and are trained on very large language datasets with additional fine-tuning on a range of NLI tasks (Vaswani et al. 2017). These models have achieved impressive performance on a wide range of natural language benchmarks, and have recently shown a remarkable ability to generate grammatically correct and relevant text in response to human queries and instructions (Ouyang et al. 2022; Chang et al. 2023; Bubeck et al. 2023). In this paper we adopt the term *vector-based* models (Blacoe and Lapata 2012) to describe methods of sentence representation in this tradition, in which a sentence is represented as a vector of numbers in a vector space without any explicitly encoded syntax.

Syntax-based approaches to sentence meaning developed from parsing methods which represent the syntactic structure of a sentence as a tree structure of nodes linked by edges. Early methods such as context-free grammars focused on specifying formal rules which determine the grammatical structure of sentences (Chomsky 1956; Kasami 1966). More recently, deep-syntactic parsing models have been developed which abstract away from much of the surface form of a sentence in an attempt to represent its underlying meaning (Kingsbury and Palmer 2002; Ballesteros et al. 2014; Michalon et al. 2016). This typically involves constructing a parse tree in which nodes are words (or other lexical items), and whose edges represent important semantic relations (e.g. predicate/argument relations) between these nodes (Žabokrtský, Zeman, and Ševčíková 2020; Donatelli and Koller 2023; Simoulin and Crabbé 2022). In this paper we use the term *syntax-based* models to describe approaches to representing sentence meaning in this tradition.[3]

In recent years, *hybrid models* that combine the complementary strengths of both syntax and vector-based approaches have been introduced(Boleda and Herbelot 2016; Ferrone and Zanzotto 2020; Donatelli and Koller 2023). Hybrid models are very diverse, and include methods for embedding parse trees into a vector representation, as well as other specialised architectures and approaches which sometimes go by the name neurocompositional semantics (Smolensky et al. 2022). What unifies hybrid approaches is a desire to integrate the distinct benefits of vector-based semantics with those of syntax-based approaches. For instance, transformers perform poorly on tasks specifically designed to test for productivity and systematicity, while syntax-based methods utilising explicit symbols easily achieve near-perfect performance (Dziri et al. 2023). Conversely, human language is highly complex and filled with nuances and idiosyncrasies, making it difficult to devise appropriate syntactic rules which describe the entirety of natural language, while vector-based methods excel at representing vagueness and nuance due to their flexibility and use of continuous numerical values rather than discrete symbols (McClelland et al. 2020). Furthermore, syntax-based methods are more readily interpretable (Linzen and Baroni 2021) and show better compositional capabilities (Yao and Koller 2022; Liang and Potts 2015), while vector-based methods excel in capturing contextual effects, integrate better with lexical semantics (Erk 2012; Pavlick 2022), and underpin existing state-of-the-art NLP applications. See Figure 1 for a visual summary of the differences between the three approaches.

---

3 While any parse tree representation can also be encoded as a vector, we classify such vector encoding of parse trees as hybrid models, since unlike traditional distributional semantics approaches they are trained to embed structured information rather than plain text. Also, unlike syntax-based approaches they collapse information into a vector-space representation, eliminating explicit representation of how semantic roles are bound to specific variables (Fodor and Pylyshyn 1988; Greff, Van Steenkiste, and Schmidhuber 2020; Vankov and Bowers 2019). We therefore believe it most useful to categorise them separately from either vector-based or syntax-based models.
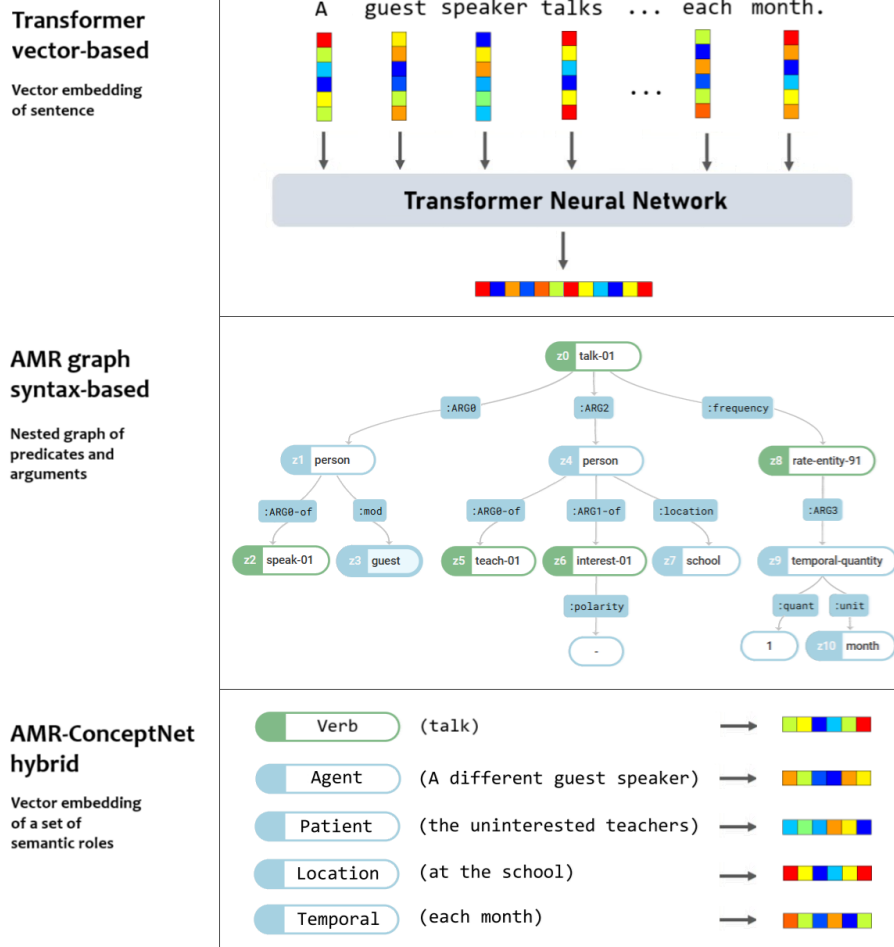
Figure 1: Illustration of three different ways of representing the sentence "A different guest speaker talks to the uninterested teachers at the school each month". Transformer neural network (top), AMR syntax-based parse tree (middle) and our novel AMR-ConceptNet hybrid model (bottom).

Despite extensive development of vector-based, syntax-based, and hybrid approaches, little work has attempted to systematically evaluate and compare how well models in each class capture human compositional representation of sentence meaning. A major difficulty is the lack of suitable frameworks and datasets for comparing such models. As we explain further in subsection 3.1, Semantic Textual Similarity (STS) provides such a measure for comparing disparate forms of sentence representation. Unfortunately, as we show in subsection 3.2, existing STS datasets based on sentential semantic similarity are inadequate for evaluating models of human compositionality. These considerations suggest the need for a novel approach to evaluate syntax, vector, and hybrid models against a common dataset.

This paper assesses how different classes of models capture human compositional representation of sentence meaning, by developing a new STS dataset. We specifically

focus on their ability to model the process of human compositionality, rather than their performance in applied tasks such as sentence parsing or machine translation. Our key contributions are twofold. First, we introduce a new dataset called STS3k, which is optimised to evaluate the strengths and weaknesses of syntax-based, vector-based, and hybrid models. Second, we evaluate several vector, syntax, and hybrid models of sentence meaning against both existing datasets and our novel STS3k dataset. Our findings show that even leading vector-based models (i.e., Transformers) poorly match human judgments of sentence similarity on our adversarial dataset, while our novel hybrid methods perform well even with no task-specific training. These key contributions together provide significant insights into formal models that describe how humans understand and represent the sentence meaning.

The remainder of this paper is structured as follows. In section 2, we review existing vector-based, syntax-based, and hybrid approaches for representing sentence meaning. In section 3, we discuss the limitations of existing STS datasets and thereby motivate our development of a new dataset. In section 4, we explain the construction of our dataset and our evaluation approach for comparing different models. In section 5, we present the results of our evaluations of the strengths and weaknesses of existing sentence models. In section 6, we discuss the implications of our results for the question of sentence representation and compositionality. In section 7, we summarise our research objectives and highlight our unique contributions.

2. Existing models of sentence meaning

In this section, we review several major approaches for representing sentence meaning. We focus on the difference between syntax-based and vector-based approaches, highlighting their distinctions, strengths and limitations. We also discuss previous attempts to integrate the two into various hybrid models, emphasising their limitations and the potential for a novel approach.

2.1 Arithmetic vector-based models

Vector-based semantics models describe word meaning as a vector of real numbers, each component of which corresponds to an abstract feature in an underlying vector space (Landauer, Foltz, and Laham 1998; Lieto, Chella, and Frixione 2017; Almeida and Xexéo 2019). The meaning of each word is thus represented by the direction of its word embedding in semantic space. Word embeddings are typically learned from statistical associations of their occurrences in large natural language corpora (Boleda 2020). They are widely used in natural language processing, either directly or as part of a machine learning pipeline, and have achieved impressive performance on a range of NLP tasks (Lenci 2018; Young et al. 2018; Devlin et al. 2019; Ranasinghe, Orăsan, and Mitkov 2019). While word embeddings capture aspects of meaning difficult to incorporate into syntactic approaches, such as vagueness and graded associations (Erk 2022), they do not come equipped with any framework for how they can be composed to form representations of an entire sentence, as this requires the specification of additional formalism beyond the word level.

An early framework for combining word embeddings into sentence embeddings was introduced by Mitchell and Lapata (2010). In this approach, the sentence embedding $\widetilde{p}$ of a given sentence is written as a function of its component word embeddings

$w_1, w_2, \ldots w_i$. This is expressed compactly by the equation:

$$\widetilde{p} = f\left(\widetilde{w}_1, \widetilde{w}_2, \ldots, \widetilde{w}_i\right)$$

The problem of representing sentences can thus be modelled as finding an appropriate function f that satisfies certain linguistic constraints and yields predictions that are psychologically plausible (Baroni 2020). Proposals for the functional forms of f are extremely diverse, with some of the most influential summarised in Table 1.

Additive/mean models are the simplest case and involve simply adding individual word embeddings component-wise to produce the sentence embedding. In some cases, the embeddings are normalised by dividing them by the number of words in the sentence, in which case the term 'mean embeddings' is used. Additive and mean models are limited because they do not incorporate any interaction effect between words, a necessity in accounting for polysemous usages such as "hot summer" compared to "hot topic" (Hartung et al. 2017). Importantly, additive models provide a useful non-compositional baseline against which more complex models can be evaluated. Two approaches for incorporating interaction effects into sentence embeddings are element-wise multiplication and circular convolution (Emerson 2020). However, both of these operations are commutative, meaning that unlike natural language, the resulting embeddings are invariant to word order (Ferrone and Zanzotto 2020). Given these limitations, more complex models have been developed. One example is the tensor product model, in which the outer product of two vectors is taken to represent the compound of those two vectors. One major drawback of such approaches is that they lead to exponentially larger embeddings for more complex expressions, as the combined embeddings scale as l^n, where l is the embedding length and n is the number of words in the expression (Stewart and Eliasmith 2009). These and other limitations of purely arithmetic models for compositional semantics have contributed to their being largely superseded by neural network models. Nonetheless, we include them in our analysis as a simple baseline for more complex models.

Table 1: Summary of arithmetic models of sentence semantics.

Model	Function	Citation
Additive/mean	$\widetilde{p} = \widetilde{s} + \widetilde{v} + \widetilde{o}$	Mitchell and Lapata (2010)
Multiplicative	$\widetilde{p} = \widetilde{s} \odot \widetilde{v} \odot \widetilde{o}$	Mitchell and Lapata (2010)
Circular convolution	$\widetilde{p} = (\widetilde{s} * \widetilde{o}) * \widetilde{v}$	Blouw et al. (2016)
Tensor product	$\widetilde{p} = (\widetilde{s} \times \widetilde{o}) \times \widetilde{v}$	Hartung et al. (2017)

2.2 Neural network vector-based architectures

More recent vector-based approaches have moved away from explicitly representing the combination function f, instead learning it implicitly by adjusting the weights in a neural network architecture in accordance with a learning objective such as next word prediction (Ferrone and Zanzotto 2020; Baroni 2020). Several architectures have been developed (Qiu et al. 2020), including recurrent neural networks (Socher et al. 2012), long-term short-term memory networks (Graves 2013), and transformers (Vaswani et al. 2017; Devlin et al. 2019). Transformers, which lack recurrent connections and rely entirely on

6

the self-attention mechanism for encoding word context, have become the most commonly used approach for sentence representation and achieve impressive performance on a wide range of language tasks (Tripathy et al. 2021; Qin et al. 2023). We summarise a selection of neural network models in Table 2. We have chosen a range of models to illustrate different architectures and training methods, including an LSTM model (InferSent), three transformer architectures optimised for producing representations of an entire sentence (USE, SentBERT, and DefSent), one general-purpose transformer optimised for text generation (ERNIE), and state-of-the-art sentence embeddings from the OpenAI API (see `https://platform.openai.com/docs/guides/embeddings`).

Table 2: Summary of neural network models of compositional semantics.

Model	Model description	Citation
InferSent	A bi-directional LSTM trained on various natural language inference tasks.	Conneau et al. (2017)
USE	Standard transformer architecture trained on a range of language tasks.	Cer et al. (2018)
SentBERT	Based on the mpnet-base transformer model, with additional training to predict paired sentences from a large dataset.	Reimers and Gurevych (2019)
ERNIE	Trained on next word prediction, masked word prediction, and prediction of hidden nodes in a knowledge graph	Sun et al. (2020)
DefSent	Based on RoBERTa-large transformer model fine-tuned using about 100,000 words paired with their dictionary definitions.	Tsukagoshi, Sasano, and Takeda (2021)
OpenAI Embeddings	Embeddings provided from the OpenAI API, based on a large transformer with additional fine-tuning from human feedback.	Ouyang et al. (2022)

Despite substantial progress, it is still an open question whether neural network models of sentence meaning provide a cognitively plausible model of sentence meaning (McCoy, Min, and Linzen 2020). Although they are able to learn aspects of sentence syntax and structure (Krasnowska-Kieraś and Wróblewska 2019; Manning et al. 2020; Pimentel et al. 2020), standard neural network architectures are not compositional in the classical sense of applying rules independently of semantic content, since combination rules are not explicitly represented, but are learned implicitly over training along with the individual word embeddings (Fodor and Pylyshyn 1988; Hupkes et al. 2020; Linzen and Baroni 2021). This violates the key criterion of compositionality that the meaning of a composite phrase is determined solely by the meaning of its constituent words and the syntactic operations for combining them (Boleda 2020). In most vector-based semantic models, structure is not defined in advance in the way that syntax is defined in formal semantics (Gajewski 2015). In theory, LLMs could learn these rules themselves, though in practice even very large models often fail to adequately and consistently generalise beyond examples found in the training distribution (McCoy, Min, and Linzen 2020; Dziri et al. 2023). Transformers are often unable to learn the types of linguistic regularities relevant to humans, instead commonly relying on lexical cues (Yu and Ettinger 2020) or spurious correlations in their training data (Geirhos et al. 2020; Niven

and Kao 2019), resulting in unsystematic and insufficient generalisation when evaluated on examples outside their training set (Hupkes et al. 2020; Gubelmann and Handschuh 2022a; Loula, Baroni, and Lake 2018; Zhang et al. 2022). A final difficulty is that in typical neural network models, the meaning of individual words is not separately represented when they are composed into a complex expression, as the network simply produces a new overall pattern of activity jointly representing all constituent words combined to form that specific sentence (Ferrone and Zanzotto 2020). Unlike traditional symbolic systems, no separate representation of individual words is preserved after composition. This makes it difficult to implement compositional rules which operate consistently across diverse examples (Fodor and Pylyshyn 1988; Martin and Doumas 2020; Mitchell and Lapata 2010).

2.3 Semantic parsing syntax-based models

Syntax-based models represent the meaning of a sentence as a graph of connected nodes, with the links between nodes reflecting syntactic or semantic relationships between components of the sentence. Individual words are typically represented in symbolic form, with formal syntactic roles governing how they can be combined together to produce valid compound expressions (Žabokrtskỳ, Zeman, and Ševčíková 2020). There is considerable variation between models in the degree of abstraction away from the surface structure of the sentence and in what types of relations are presented. This variation takes the form of different grammars, the sets of formal rules specifying how nodes in the resulting graph are combined (Zhang 2020). Examples of some major contemporary frameworks are presented along with brief explanations in Table 3. These all share a common approach of first identifying key verbs or predicates, and then associating various semantic roles to those predicates. The set of semantic roles is usually predetermined based on linguistic theory, and may be constant for every predicate or different for each one. Some frameworks (such as AMR and UCCA) also provide a nested structure of relations between sentence components, while others (such as FrameNet and VerbNet) only specify the relation between the main predicate and its arguments at a single layer without any nested structure. While we highlight a range of different approaches to illustrate the range of formalisms that have been developed, in this paper we selected for further analysis VerbNet (based on semantic role labelling) and AMR (a graph-based method), owing to their flexibility and the availability of efficient parsing algorithms.

Syntax-based methods for semantic parsing have been the focus of much theoretical work in semantics, and recently have seen an increase in attention due to the development of more sophisticated neural network parsing algorithms and the availability of much larger annotated datasets (Bölücü, Can, and Artuner 2023). Because they describe the logical connections between different components of a sentence in a readily extensible manner that separates variables from their values, syntax methods readily support compositional reasoning, at least for constrained problems. On the other hand, these methods typically treat individual words as undefined primitive symbols and thus provide no clear interface between lexical semantics and compositional semantics (Erk 2016). Manual parsing rules often fail to represent language variability and phenomena such as polysemy or connotation. A further challenge is the difficulty in evaluating syntax-based models using a similarity metric analogous to the cosine similarity widely used for assessing vector-based embeddings. The SMATCH metric (Cai and Knight 2013), along with SMATCH-based variations like WWLK (Opitz, Daza, and Frank 2021a), are widely-used for computing the similarity of two parse graphs.

Table 3: Summary of major syntax-based approaches to representing sentence meaning.

Model	Description	Citation
PropBank	A corpus and annotation framework based around verbs and their arguments, with generic argument roles applied to each verb.	Kingsbury and Palmer (2002)
VerbNet	An annotation and classification scheme for verbs, incorporating a standardised set of thematic roles and selectional preferences depending on the verb.	Palmer, Bonial, and Hwang (2016)
FrameNet	A database of lexical frames, each of which describes a particular type of event or relation and the elements that participate in it.	Baker, Fillmore, and Lowe (1998)
AMR	*Abstract Meaning Representation* is a graph-based framework rooted at the main verb of a sentence. Verb arguments are assigned to nested components of the sentence.	Banarescu et al. (2013)
UCCA	*Universal Conceptual Cognitive Annotation* is a graph-based approach to represent sentence meaning in terms of key abstract nodes and a determined set of relations between them.	Abend and Rappoport (2013)

However, recent studies have found that the results show a very low correlation with human similarity judgements (Leung, Wein, and Schneider 2022). We discuss this issue in more detail in subsection 4.4.

2.4 Hybrid approaches

The fact that syntactic and vector-based semantic models have complementary strengths and weaknesses has led to considerable interest in combining these approaches (Padó and Lapata 2007; Boleda and Herbelot 2016; Ferrone and Zanzotto 2020; Martin and Baggio 2020). Although standard transformer architectures have been shown to learn some aspects of sentence structure and semantic relations implicitly, such learning is still imperfect and is likely to be inadequate for robust, comprehensive sentence representations (Zhang et al. 2020a; Hupkes et al. 2020). As such, the goal of much recent work has typically been to augment transformers with explicit information about syntactic relations and semantic roles (Colon-Hernandez et al. 2021; Bai et al. 2021). The most common method is to inject such information during training using treebanks or other syntactic data (Yu et al. 2022). Several recent approaches to such hybrid models are summarised in Table 4. While we include a range of models in the table to highlight the diverse range of approaches, we selected S3BERT and AMRBART as a representative hybrid models for further analysis, as they utilise information from AMR graphs, thereby providing a useful comparison to other AMR-based methods we analyse.

An examination of recent hybrid models highlights several challenges. First, the range of approaches is extremely broad, with little consistency between them and

often minimal theoretical justification of each method (Colon-Hernandez et al. 2021). This makes interpretation of results difficult, especially since even small variations in preprocessing can significantly impact parsing performance (Kabbach, Ribeyre, and Herbelot 2018). Second, as shown in the 'increase in correlation' column of Table 4, none of these approaches substantially improve their ability to describe human judgements of sentence similarity, with most models only achieving a 1-2 percentage point increase in correlation against STS datasets relative to traditional vector-based models. Third, Yu et al. (2022) recently showed that augmenting transformers with entirely uninformative parse graphs can improve their performance on various benchmarks, in line with previous results for Tree-LSTMs (Shi et al. 2018), suggesting that these improvements may be due to a greater depth of processing of existing input rather than any crucial role of syntactic information as such.

Table 4: Summary of hybrid models of sentence meaning. The column 'increase in correl.' shows the percentage point increase in correlation over the best-performing comparable non-hybrid model (e.g. BERT), as reported in the original paper.

Model	Explanation	Increase in correl.	Citation
DRS	Sentence similarity computed as a weighted average of word order, constituency parse, and embedding similarities.	1.30	Farouk (2020)
SemBERT	PropBank semantic roles extracted and encoded into vectors using BERT. These roles are concatenated with word embeddings to produce sentence embeddings.	0.20	Zhang et al. (2020b)
Syntax-BERT	Mask matrices computed with semantic parsers indicate which words are syntactically connected. Transformer attention was then augmented with these mask matrices.	2.00	Bai et al. (2021)
SynWMD	Syntactic distance between components is estimated by dependency parsing. Sentence similarity is then computed by word mover distance weighted by syntactic distance.	0.84	Wei, Wang, and Kuo (2023)
AMRBART	A method based on the BART transformer for embedding an AMR graph into a vector.	-	Bai, Chen, and Zhang (2022)
S3BERT	Modification of SentenceBERT to incorporate information from AMR parsing of sentences. Also decomposes the sentence similarity score into constituent AMR features.	0.60	Opitz and Frank (2022a)
EF-SBERT	Constituency-parsed semantic elements are passed through a transformer, then combined with a full sentence embedding.	0.54	Wang et al. (2022)
SpeBERT	Words are paired using dependency parsing to compute part embeddings, which are concatenated to give full sentence embeddings.	1.92	Liu et al. (2023)

Given these difficulties, we have developed an alternative approach to develop novel hybrid models. Instead of attempting to inject information about semantic roles and syntax into transformers, we take individual word embeddings and then combine them in accordance with the sentence structure or semantic roles specified by a syntax-based method. This effectively means using vector-based models at the level of lexical semantics and syntax-based methods at the level of compositional semantics. The aim is to combine the flexibility and gradedness of vector-based embeddings with the explicit structure of syntax-based methods. We explain our novel approach in more detail in subsection 4.4.

3. Testing models of sentence meaning

Having presented an overview of existing models for representing sentence meaning, we now consider different methods for evaluating such models. Given our interest is in assessing how accurately different models of sentence representation describe the cognitive mechanisms for sentence representation in humans, we focus on evaluation methods capable of testing the *representations* (graphs or embeddings) of different formalisms rather than their performance on downstream tasks. We are thus interested in the cognitive plausibility of these models - their ability to form human-like representations of sentences, not just whether they are able to perform well in language tasks. As such, in this section we review the Semantic Textual Similarity approach to evaluation, outline important limitations of existing datasets and the need for better data, and place our work in the context of other approaches to evaluating compositionality in models of sentence meaning.

3.1 Semantic textual similarity

Semantic Textual Similarity (STS) involves collecting human judgements of semantic relatedness or similarity for sets of sentence pairs. A model is assessed against an STS dataset by computing the cosine similarity of the embeddings assigned to each sentence and then calculating the correlation with human judgments, with higher values indicating a better performance (Erk 2012; Amigó et al. 2022). STS is an established and widely-used method for evaluating models of sentence representation (Mitchell and Lapata 2010; Krasnowska-Kieraś and Wróblewska 2019). In contrast to evaluations using downstream performance, the STS task provides a more direct assessment of the structure of the model representations (Bakarov 2018; Pavlick 2022), which makes it easier to identify which aspects of the model are beneficial or detrimental (Ribeiro et al. 2020a; Bakarov 2018; Pavlick 2022). We consider various criticisms of STS as an evaluation method in subsubsection 8.1.3 of the Appendix.

The major English STS datasets are summarised in Table 5. The three largest datasets, STSb, SICK, and STR-2022, are constructed from sentences extracted from various online sources, mainly news headlines, forum posts, image captions, Twitter, book reviews, and video descriptions. The smaller STSS-131 dataset consists of sentences between ten and twenty words long, all constructed from dictionary definitions. The GS2011 and KS2013 datasets consist of simple subject-verb-object sentences originally developed to test models of categorical compositional grammar. All are annotated by crowd-sourced participants without specific training, though the precise instructions vary across the datasets. Several other less directly relevant datasets are discussed in subsubsection 8.1.2 of the Appendix.

Table 5: Summary of English STS datasets.

Dataset	Stimuli	Type	Raters	Citation
STSb	8,628	Sentence pairs	5	Agirre et al. (2016)
SICK	10,000	Sentence pairs	10	Marelli et al. (2014)
GS2011	200	Sentence pairs	6	Grefenstette and Sadrzadeh (2011)
KS2013	108	Sentence pairs	24	Kartsaklis, Sadrzadeh, and Pulman (2013)
STSS-131	131	Sentence pairs	64	O'shea, Bandar, and Crockett (2014)
STR-2022	5,500	Sentence pairs	8	Abdalla, Vishnubhotla, and Mohammad (2023)
Mitchell-324	324	Bigram pairs	18	Mitchell and Lapata (2010)
BiRD	3,345	Bigram pairs	17	Asaadi, Mohammad, and Kiritchenko (2019)
STS3k	2,800	Sentence pairs	20	This paper

3.2 The need for better datasets

Despite the value of the STS approach, existing STS datasets using natural language sentences have significant limitations. First, the datasets are not optimized for comparing multiple models in terms of how well each model captures human judgment similarity. As we show in subsection 5.2 and Figure 3, even entirely non-compositional models score high correlations against them. A recent small-scale study of fifty complex sentences showed a similar result in the STSb and SICK tasks (Chandrasekaran and Mago 2021). These results indicate that the human ratings in existing datasets primarily reflect the degree of lexical similarity of the words in each sentence, rather than the degree of similarity of sentence structure. In other words, existing datasets exhibit a major confound between lexical and structural similarity, which makes it difficult to assess the adequacy of different models of sentence meaning since even simple non-compositional models perform about as well as much more sophisticated models.

Furthermore, the data are of highly variable quality and in the case of STSb, SICK, and STR-2022, include many sentence pairs that are ambiguous, ungrammatical, too simplistic, or too complex to be ideal for testing models of compositional semantics (see Table 6 for examples). This stems from the automated selection of sentences from online forums, tweets, headlines, and image captions. Many headlines and image captions are not grammatical sentences, which make parsing difficult and poorly serves the objective of testing representations of sentence meaning. Others lack sufficient context for humans to adequately judge their meaning. Many sentences are also either very short (less than about five words) or very lengthy and convoluted (more than twenty words with multiple clauses).

These limitations highlight the need for a new STS dataset with a more carefully curated set of sentence pairs designed specifically to facilitate comparisons between different representations of sentence meaning. Stimuli need to be carefully designed

Table 6: Summary of problems with existing STS datasets. Examples have been chosen to illustrate many similar sentences found in the datasets STSb, SICK, and STR-2022.

Issue	Example sentences	Comments	Frequency
Simplistic structure	- A dog is barking. - A man is playing a violin. - A man is frying a tortilla.	Sentences with only a copula verb are often too simple to assess composition.	In image caption portion, 3218 of 6500 sentences contain only the verb 'is'.
Sentence fragments, lacking in context	- You should prime it first. - How do you do that? - 5 nations meet on haze. - Well I wouldn't risk it, not in a cold compost system. - Websites battle nasty comments, anonymity.	Human judges likely to struggle with missing words or lack of context to assess meaning.	In a sample of 50 sentences from the deft-forum portion (forum posts), we found 13 (26%) have undefined pronouns or are sentence fragments.
Very short sentences	- People walk home. - A man is talking. - The gate is blue.	Too simple to be useful for assessing meaning composition.	Of 3000 sentences in the MSRvid portion (video captions), 256 have only four words or fewer.
Very long sentences	The Justice Department filed suit Thursday against the state of Mississippi for failing to end what federal officials call 'disturbing' abuse of juveniles and unconscionable conditions at two state-run facilities.	Too complex to be ideal for assessing meaning composition.	Common in the MSRpar portion (news headlines), where 903 of 3000 sentences are longer than 20 words.
Unfamiliar acroynms, proper nouns	- Results from No. 2 U.S. soft drink maker PepsiCo Inc. (nyse: PEP - news - people) were likely to be in the spotlight. - Serrano * ES 4705 D m (2)	Sentences cannot be understood without prior knowledge that humans and models may lack.	Common in MSRpar portion (news headlines), with an average of 4.3 uppercase letters per sentence, owing to many acronyms and proper nouns.

to ensure that only models sensitive to sentence structure will score high correlations against the dataset, thereby controlling for the confound of lexical similarity. Furthermore, all stimuli should consist of complete grammatical sentences with sufficient context for raters to properly understand their meaning. It is also important to strike the right balance between sentences that are sufficiently complex to contain variable structure that affects overall meaning, without being so complex that they are difficult for human raters or parsing algorithms to assess. These principles inform the design choices we made in constructing our novel STS3k dataset, as described in subsection 4.1.

In this study, we attempt to overcome the weaknesses of existing STS datasets by developing a novel dataset called STS3k, consisting of 2,800 sentence pairs rated for

semantic similarity by human participants. Using our novel dataset, we compare vector-based (including non-compositional baseline models and transformer neural networks), syntax-based, and hybrid models in terms of how well each model captures human judgments of sentence similarity.

3.3 Compositional tasks

Our novel STS dataset also draws inspiration from an alternative approach to evaluating sentence representations using compositional generalisation tasks. The focus of these tasks is to examine how language models learn the underlying structure of a textual input to perform out-of-distribution generalisation. The importance of structure and compositional generalisation was one of the guiding principles in the construction of our task. We discuss this issue in more detail in subsection 8.1 of the Appendix.

4. Methods

4.1 Dataset construction

Here we introduce our novel dataset STS3k[4]. It consists of 2,800 handcrafted sentence pairs rated for semantic similarity by crowd-sourced respondents. In order to avoid some of the limitations described in subsection 3.2, we developed a set of sentences adhering to a specified structure, designed to test specific aspects of compositional models. The motivation behind developing our dataset was to develop a new benchmark that combines the systematic variation in components of the compositional tests with the linguistic plausibility and structure of natural language sentences from STS datasets. To provide a more controlled set of stimuli for testing the impact of sentence structure on similarity, and to mitigate some of the limitations of existing STS datasets described in subsection 3.2, our dataset includes only single-clause sentences consisting of a subject, a verb, a direct object, and various combinations of optional elements. Sentences follow the following structure, with optional elements (which typically correspond to adjuncts for the main verb) shown in square brackets:

$$\text{Sentence} = [\text{Adjective}] + \text{Subject} + [\text{Adverb}] + \text{Verb} + [\text{Adjective}] + \text{Object}$$
$$+ [\text{Manner}] + [\text{Adjective}] + [\text{Indirect Object}] + [\text{Time}] + [\text{Place}] \tag{1}$$

Sentences were constructed excluding the following syntactic dependencies and word types:

- *Auxiliary verbs, including modal verbs*. This was to ensure each sentence had only a single verb. An exception was made for sentences converted to passive voice.

- *Conjunctions*. These are unnecessary as sentences consist of a single declarative clause.

- *Pronouns*. These words convey little semantic content and were replaced with an appropriate regular noun.

---

4 The STS3k dataset and related code is available at
https://github.com/bmmlab/compositional-semantics-eval

- *Proper noun*s. These have semantic properties different from regular nouns and may be unfamiliar to some participants.

- *Explicit negation*. Negation is especially difficult to encode in vector-based models, and we decided to leave this aspect to further research.

Sentences pairs were designed to systematically vary different semantic elements to test the effect of each element on the meaning of the overall sentence. The different sentence types are described in detail in Table 7. The dataset consists of two portions: non-adversarial and adversarial. The *non-adversarial* portion is comparable to existing STS datasets (though with a more controlled structure), consisting of sentences with varying numbers of components changed. It serves as a baseline of comparison for the adversarial portion. For example, one non-adversarial sentence pair is:

Malaria infects millions of people each year.

Malaria occurs mostly in tropical countries.

The *adversarial* portion of the dataset is inspired by adversarial approaches to machine learning models, where a set of stimuli is constructed deliberately to probe the capabilities of a particular model or technique (Nie et al. 2020). We are interested in developing a dataset on which entirely non-compositional models perform poorly, thereby allowing us to test more directly for compositional capability and avoid the limitation of existing STS datasets discussed above, on which even non-compositional methods perform well. The key consideration was therefore to generate sentence pairs where lexical similarity was dissociated with overall similarity in meaning. This takes two forms: sentence pairs with *low* lexical similarity but relatively *high* similarity in overall meaning, or with *high* lexical similarity but relatively *low* similarity in overall meaning.

To achieve the first case (low lexical similarity but high overall similarity), we developed two types of sentence pairs: 'paraphrases' and 'added modifiers'. Paraphrase pairs contain two sentences designed to have a very similar overall meaning with minimal lexical overlap. Added modifier pairs were constructed by keeping the same major sentence roles (subject, verb, and direct object) fixed, while adding various numbers of secondary modifying elements such as time, manner, place, trajectory, or adjectives, thereby reducing lexical overlap while keeping the core meaning similar across both sentences. For example:

The plane crashed in the desert.

The cargo plane crashed in the rocky desert near the oasis at night.

To achieve the second case (high lexical similarity but lower overall similarity), we developed three types of sentence pairs which we call 'swaps'. These involve interchanging two or more words within a sentence, leaving the transformed sentence with (mostly) the same words as before but with the words now serving different roles. For

example:

<div align="center">

The professor asked the student for help.

The student asked the professor for help.

</div>

Here the subject and direct object ("professor" and "student") have been interchanged, yielding a sentence with the same words but a different meaning. Because two elements have been interchanged, we call this a 'double swap'. An example of an even more strongly adversarial pair is:

<div align="center">

The firm paid for the project with the new government.

The new government projected increased pay for the firm.

</div>

Here there have been four interchanges of word components: the subject and indirect object ("firm" and "new government") and the verb and direct objects ("paid" and "project") in this case, with some minor modifications to ensure grammatically. Since four elements have been interchanged, we call this a 'quadruple swap'.

The adversarial portion of our dataset consists largely of variations of this approach of interchanging different sentence components. This ensures that entirely non-compositional models, such as mean word embeddings, will give high similarity ratings to such sentences because they contain mostly the same words. Only models that correctly identify the structure of the sentence and the relationship between the different components are expected to yield accurate similarities. In addition, we also include some 'negative' sentences which have mostly the same words, but express the opposite meaning due to implicit negation.

As we show in subsection 5.2, particularly Figure 3, this method of construction succeeded in generating a dataset which differentiates between compositional and non-compositional models of sentence meaning by removing the confound of lexical similarity.

Table 7: Summary of types of sentence pairs included in the STS3k dataset.

Pair Type	Count	Adversarial	Explanation
Zero	70	No	Two sentences with no words in common and no obvious similarity in meaning.
Adjective	61	No	A single adjective added before the subject, direct object, or indirect object in one of the sentences.
Constant verb	96	No	Verb is kept the same, but all other components are changed.
Constant dobj	88	No	Direct object is kept the same, but all other components are changed.
Constant subj	94	No	Subject is kept the same, but all other components are changed.
Single change (verb)	137	No	Subject and direct object are kept the same, but verb and modifiers are changed.
Single change (dobj)	138	No	Subject and verb are kept the same, but direct object and modifiers are changed.
Single change (subj)	153	No	Direct object and verb are kept the same, but subject and modifiers are changed.
Other	218	No	Variants that do not fit into the above categories, mostly involving ad hoc interchanges of various sentence elements.
Check	10	No	Attention check items.
Paraphrase	71	Yes	Two sentences with similar meanings but few or no words in common.
Added modifiers	679	Yes	Two sentences with the same major semantic roles (subject, verb, and direct object), but with between one and six modifiers added to one sentence in the pair.
Double swap	538	Yes	Either the verb and the direct object, or the verb and the subject, or the direct object and the subject are swapped, leaving the third element unchanged.
Triple swap	197	Yes	All three of the verb, direct object, and subject are interchanged.
Quadruple swap	179	Yes	All four of the verb, direct object, indirect object, and subject are interchanged.
Negative	71	Yes	Two sentences which describe opposite situations, but without using explicit negation words like 'not'.
Total	2800	1735	

4.2 Human similarity judgments

A total of 523 participants (322 male, 167 female, and 12 other; age range, 18-45; mean age $\pm$ SD, 32.4 $\pm$ 7.0) were recruited using the Prolific platform (https://www.prolific.com/). Participants were paid £4.50 for completing the task, which took an average of 24.6 minutes, amounting to an hourly rate of £10.96. All participants were self-declared native English speakers in Australia or the United States. The study protocol was approved by the University of Melbourne Human Research Ethics Committee (Reference Number: 2023-23559-36378-6).

Each participant provided similarity judgments on a 7-point Likert scale (1-7) of 110 sentence pairs randomly selected from the pool of 2,800 pairs. Given the inherent vagueness of the similarity judgement task, previous studies have noted that detailed instructions on how to make similarity judgements are often unclear, or may bias participant responses (Abe et al. 2022; Abdalla, Vishnubhotla, and Mohammad 2023). Because our goal was to elicit intuitive judgements without imposing any particular framework which might bias results towards a subset of models, we did not provide participants with any special training or instructions about how to allocate ratings. We simply asked them to 'consider both the similarity in meaning of the individual words contained in the sentences, as well as the similarity of the overall idea or meaning expressed by the sentences'. The full instructions given to participants can be found in subsubsection 8.2.1 in the Appendix.

Participants were also presented with additional 10 sentence pairs that served as an attention check. These stimuli consisted of either pairs of identical sentences (high similarity) or one simple sentence paired with a grammatically correct but nonsensical sentence (low similarity). We excluded all participants who failed more than one of the attention check items, resulting in 501 out of 523 participants being retained. This amounted to 55,110 judgments, providing an average of 20 ratings for each sentence pair. Similarity judgements were averaged over participants and normalised between 0 and 1 to yield the final STS3k dataset.

4.3 Evaluation of vector-based models

We evaluated various vector-based models, including non-compositional Mean, Mult (multiplication), and Conv (convolution) models, as well as all the neural network models (see Table 2 for details), based on the consistency between the model-predicted similarities of sentence pairs and human judgments of the similarity of sentence pairs. To obtain the model-predicted similarities on the STS3k dataset, cosine similarities of the sentence embedding vectors between sentence pairs were computed. The cosine similarities were then compared with human similarity judgments using the Spearman correlation coefficient to evaluate model performance. We utilise the Spearman rank correlation since different models may give different distributions of similarities (e.g. some may tend to rate most sentences high, others may tend to rate them low), the Spearman correlation coefficient considers only the relative ordering of sentence similarities, which can be meaningfully compared across models.

Sentence embeddings using the Mean, Mult, and Conv methods were computed by performing the corresponding operation element-wise on the ConceptNet word embeddings for each word of the target sentences after removing stop words. All other vector-based models (including InferSent and all transformer-based models) were utilised as pre-trained models without any further modification or training. The last output layer was used for neural network architectures designed specifically for rep-

resenting sentences (InferSent, USE, SentBERT, DefSent, and OpenAI Embeddings). In the case of the general-purpose transformer ERNIE, we computed cosine similarities using both the input layer (layer 0) and the final layer (layer 12). For all transformers, the sentence embeddings were normalised by subtracting the mean and dividing by the standard deviation of each feature. This was found to improve the correlation with human judgements, and is motivated by previous research indicating that without normalisation, transformers tend to learn very anisotropic embeddings with a few dimensions dominating over all the others (Timkey and van Schijndel 2021; Cai et al. 2021). See Table 12 in the Appendix for details of all the models tested.

4.4 Evaluation of syntax-based models

We adopted AMR as a representative syntax-based model for representing sentence meaning. We used the SapienzaNLP (Spring) AMR parser (Bevilacqua, Blloshmi, and Navigli 2021) to parse all sentences, as it is among the best-performing AMR parses with freely available and easily implementable code. As discussed in subsection 2.4, evaluating syntax-based models also requires a method for computing the similarity between the graphs for each sentence. While various techniques have been developed for converting graphs into vector embeddings, these have typically focused on knowledge databanks rather than natural language (Goyal and Ferrara 2018; Rossi et al. 2021). Furthermore, we are interested in testing graph-based models of representing sentences more directly, rather than the embeddings produced from these graphs. As such, we analyse the similarity of AMR-embeddings using two several existing methods for comparing graph similarity directly: SMATCH (Cai and Knight 2013) and WWLK (Opitz, Daza, and Frank 2021a). The corresponding sets of similarity ratings are therefore referred to as AMR-SMATCH and AMR-WWLK, indicating both the deep-syntactic formalism used and the method of similarity adopted for comparing sentences. As for the vector-based models, the fit to human similarity judgments was estimated by computing the Spearman correlation coefficient between model similarities and human judgements.

4.5 Evaluation of existing hybrid models

Two of the hybrid models we examine (AMRBART and S3BERT) utilise vector embeddings for the final sentence representation, and so for these models use cosine similarity to compute sentence similarities. Once again, the fit to human judgements was estimated using the Spearman correlation coefficient.

4.6 Development of novel hybrid models

Inspired by previous work (Salehi, Cook, and Baldwin 2015; Wang, Mi, and Ittycheriah 2016; Farouk 2020), we developed a novel method for evaluating the similarity of parse trees as a linear combination of the similarity of various sentence components. The key idea is to represent each sentence as a combination of the major semantic roles that describe the relevant situation (Chersoni et al. 2019). Several previous studies have implemented comparable hybrid methods using weighted averages of various sentence elements. Farouk (2020) compute sentence similarity as the weighted average of word order similarity, constituency parse similarity and overall embedding similarity. Wang et al. (2022) and Liu et al. (2023) each implement a slightly different method involving computing embeddings of constituency-parsed semantic elements of a sentence, which

are then combined together to produce a full sentence embedding. Our approach is designed to combine the flexibility and gradedness of vector-based models with the explicit structure provided by syntax-based models. The major downside to this method is that we are only able to incorporate specific predefined aspects of syntax. We discuss this in further detail in subsection 6.3.

Our two novel hybrid models differ from previous hybrid approaches in that they do not use any neural network architecture at all, nor do they represent a sentence using a single final embedding. Instead, the meaning of a sentence is represented as a *set* of embeddings (see Figure 1 for an illustration), each of which is computed by averaging the word embeddings for all words in a given parse element, where elements are taken from either AMR (Banarescu et al. 2013) or VerbNet (Palmer, Bonial, and Hwang 2016) depending on the model. In the VerbNet case, we parsed each sentence using a VerbNet semantic role labelling algorithm, then computed the embeddings for each role by averaging over the static ConceptNet word embeddings for each word associated with that role. As far as we are aware, use of ConceptNet word embeddings for such a purpose is also novel. These were chosen as the highest performing word embeddings on word similarity datasets (Fodor, De Deyne, and Suzuki 2023). The overall sentence similarity was then computed as the weighted average of role-wise similarities. In the AMR case, we first parsed each sentence using an AMR parser, then computed the embeddings for each level of the parse tree by averaging over static ConceptNet word embeddings of each leaf node. Leaves at the same level of the parse tree are then aligned, and the overall sentence similarity is computed as the weighted average of these aligned leaf embeddings. We refer to these hybrid models as "AMR-CN" and "VerbNet-CN" to emphasise that they involve combining the relevant parsing method with the ConceptNet (CN) word embeddings. Below we outline our process for computing the similarity of VerbNet semantic role and AMR parses of sentence pairs in detail.

VerbNet-CN similarities were computed as follows:

1. Compute the VerbNet semantic roles for each sentence using the SemParse Docker image provided in the SemLink project (Gung 2020). We used this as a high-performing and easy-to-use semantic role labelling algorithm.

2. Manually adjust the automated output to ensure consistency and rectify improperly parsed sentences. Improper parsing was usually the result of failing to correctly identify the main verb of the sentence or inconsistently classifying similar elements into different roles.

3. Consolidate all semantic roles into eight basic categories. These were based on the General Thematic Roles from the VerbNet Unified Verb Index[5]. In addition to the Verb, we selected the most commonly used roles Agent, Patient, Theme, and grouped most of the less common roles into Location, Trajectory, Manner, and Place. As an additional check on our method, we used the GPT-4 model of the OpenAI Chat Completions API[6] to directly parse each of the STS3k sentences using the same eight semantic roles. We give the full instruction in subsubsection 8.2.1

---

5 See documentation at https://uvi.colorado.edu/references_page
6 https://platform.openai.com/docs/guides/text-generation/chat-completions-api

4. Compute the embeddings of each semantic role by averaging the static ConceptNet embeddings of each constituent word after the removal of stop words. Words that are not associated with any semantic role are discarded.

5. Compute the cosine similarity between the embeddings of each semantic role of the first sentence with the corresponding semantic role of the second sentence. This yields eight similarity scores, which we refer to as the RoleSims, one for each semantic role.

6. To improve matching between similar sentences with different structures, we paired non-identical semantic roles when no exact match could be found. For example, if one sentence had an Agent but no Patient, while the second sentence had a Patient but no Agent, then the Patient and Agent similarity would be calculated and used in the calculation for the overall sentence similarity. This matching process was hard-coded to operate in the same way for all sentence pairs.

7. Finally, compute the sentence similarity as a weighted average of RoleSims. This is depicted in Equation 2, where s_1 and s_2 represent the two sentences to be compared, and $r_{i,1}$ and $r_{i,2}$ represent the semantic role embeddings for role i.

8. The RoleSim weights β_i were chosen using a separate pilot dataset consisting of simple subject, verb, and object sentence pairs along with similarity ratings provided by human participants. By rounding the estimated parameters from this pilot data, we set the weight of 3 for the Verb and 2 for Agent, Patient, and Theme. As this pilot data only included simple sentences without additional semantic roles, we selected lower weights of 0.5 for Time, Manner, Location, and Trajectory based on the intuition they would have less impact on sentence meaning than Agent, Patient, or Theme. We opted to use fixed parameters rather than learn them from the STS3k data to avoid giving the VerbNet-CN model an unfair advantage over the transformer models, which had no parameters adjusted based on the STS3k dataset. Also, as shown in Figure 6, the performance of VerbNet-CN is not dramatically changed even when the parameters are learned directly from the STS3k dataset. All pilot data is available on our GitHub repository.

$$\text{SentSim}\,(s_1, s_2) = \sum_{i=1}^{8} \beta_i \cdot \text{RoleSim}\,(r_{i,1}, r_{i,2}) \tag{2}$$

AMR-CN sentence similarities were computed as follows:

1. Sentences were parsed using the SapienzaNLP (Spring) AMR parser (Bevilacqua, Blloshmi, and Navigli 2021).

2. Each token in the sentence was assigned an 'AMR role' in accordance with its location in the parse tree. This was constructed by concatenating all nested parse labels.

Table 8: Parameters used for training a feed-forward neural network for fine-tuning against the STS3k dataset.

Num hidden layers	Hidden units	Total parameters
0	0	8
1	10	80
2	60, 100	1,090
3	100, 100, 10	11,810

3. Role similarities were computed as the cosine similarity between the averaged ConceptNet word embeddings for all tokens with the same AMR role in each sentence of a sentence pair.

4. Compute the final sentence similarity as average role similarity over all roles found in either sentence:

$$\text{SentSim}\,(s_1, s_2) = \frac{1}{n} \sum_{i=1}^{n} \text{RoleSim}\,(r_{i,1}, r_{i,2}) \tag{3}$$

4.7 Fine-tuning against the STS3k dataset

To investigate whether fine-tuning against our STS3k dataset would improve model performance, we developed a series of models to predict human similarity judgements by training a classifier using the STS3k dataset. Following a similar methodology to that used in previous studies (Reimers and Gurevych 2019; Etcheverry and Wonsever 2019), we trained a simple classifier taking the concatenated embeddings for two sentences as input and outputting a number between 0 and 1, corresponding to the predicted human-rated similarity of the two sentences. Because our VerbNet-CN model has only eight parameters (one for each of the semantic roles), we first used principal component analysis to reduce the dimensionality of the sentence embeddings for each arithmetic and neural network model, retaining the top eight components to match the number from VerbNet-CN. We then trained simple feed-forward neural networks with between zero and three hidden layers, each fitted using a subset of the STS3k dataset and evaluated on a holdout testing subset. The number of hidden units and total number of parameters is shown in Table 8. We selected the number of hidden units so that the total number of parameters increased by roughly a factor of ten for each additional layer. The models were trained using sklearn MLPRegressor 1.2.2 with default parameters. We trained two sets of models, first a random train/testing split and then a split where the model was trained on the non-adversarial subset and tested on the adversarial subset (excluding negatives). The purpose of the latter split was to analyse out-of-distribution generalisation, a crucial component of compositional reasoning. As an additional check, we also performed this analysis without any dimensionality reduction.

In a separate analysis, we performed a full fine-tuning of the SentenceBERT model using a script provided by the authors of this model (Reimers 2021). All parameters

in the model were adjusted during training over 1,000 evaluation steps and 4 training epochs. As before, we performed the fine-tuning using a random train/testing split, and also a split based on training on the non-adversarial subset of STS3k and testing on the adversarial subset.

5. Results

In this section, we begin by presenting key descriptive statistics and assessing the quality of our STS3k dataset. We then use our dataset to evaluate a range of models of sentence meaning, first without any specific training on our dataset, and then with fine-tuning on the STS3k dataset. Finally, we investigate the STS3k results in more depth to determine what effect different sentence components have on human judgements of sentence meaning.

5.1 STS3k dataset descriptive statistics

The normalised sentence similarity ratings ranged from 0 to 0.975 with the $mean = 0.442$ and $SD = 0.242$. As shown in Figure 2, the shape of the ratings histogram is significantly different from that obtained by randomly shuffled ratings ($p = 4.3 \times 10^{-119}$, *Kolmogorov–Smirnov* test). Those results indicate that the similarity ratings cover almost the entire range of the rating scale in a systematic non-random manner.

To assess the consistency of ratings across participants, we computed the average standard deviation of similarity scores for each sentence pair across participants. We found this to be equal to 0.216, which is comparable to the 0.19 adjusted average standard deviation of the SICK dataset (Marelli et al. 2014) and slightly above the 0.163 of the STSS-131 dataset (O'shea, Bandar, and Crockett 2014). Moreover, we computed
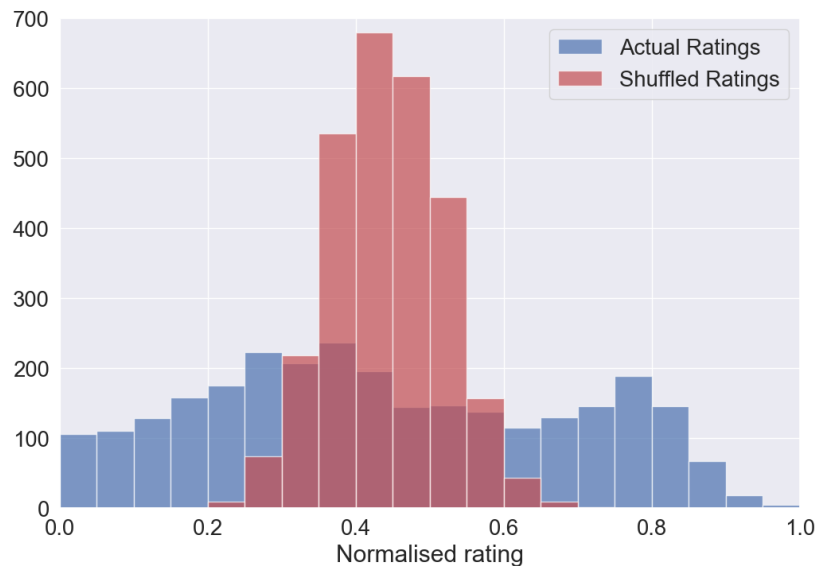

Figure 2: Histogram of normalised ratings of sentence pairs, showing the actual distribution and the distribution obtained by shuffling ratings within each participant.

the split-half reliability with the Spearman-Brown correction for the entire dataset as 0.953, indicating very high agreement between individual raters. Inter-rater agreement was also very high for each portion of the dataset, with values of 0.950 for the non-adversarial and 0.940 for the adversarial portions respectively. We also computed linearly weighted Cohen's kappa using the same split-half method, finding values of 0.832 for the entire dataset, 0.825 for the non-adversarial portion, and 0.804 for the adversarial portion.

5.2 Comparative evaluation of sentence models

We next evaluated the fit of each computational sentence model with existing STS datasets and our new STS3k dataset *without* any additional training. For this purpose, we computed the Spearman correlation between model-derived similarities and human similarity ratings of sentence pairs. The complete set of results is shown in Table 9. As described in subsection 4.3, sentence similarities for all vector-based models were computed using cosine similarity. Sentence similarities for the syntax-based models are computed using SMATCH or WWLK metrics, as outlined in subsection 4.4 . Similarities for the novel hybrid methods (AMR-CN and VerbNet-CN[7]) were computed as described in subsection 4.6.

Existing datasets cannot effectively discriminate between sentence models. We first demonstrate that existing datasets cannot differentiate between the non-compositional Mean-CN model and more complex models of interest. On the existing STSb, SICK, and STSS-131 datasets, even the non-compositional Mean-CN model performs fairly well compared to other neural-network, syntax-based, and hybrid models (Table 9), indicating that current STS datasets cannot effectively discriminate between these models. For example, as illustrated in Figure 3, there is only a small difference in the Spearman correlation (0.1 − 0.2 points) between the non-compositional Mean-CN model and the DefSent transformer neural network model. Similar levels of difference (0.1 − 0.3 depending on the dataset) are observed for other neural network models, including OpenAI and SentBERT. The Spearman correlations of syntax and hybrid models are highly variable, ranging from even lower than the non-compositional Mean-CN model in the case of AMR-WWLK, to being comparable to the best transformer models in the case of S3BERT. Overall, these results indicate that existing datasets are easy even for entirely non-compositional models, and hence are inadequate for testing models of human representation of sentence meaning.

Our STS3k dataset can discriminate between sentence models. By contrast, on our new STS3k dataset, the gap between the non-compositional Mean-CN model and other complex models is much larger. This difference is best illustrated by comparing the non-adversarial portion of the STS3k dataset to the adversarial portion (see Figure 3). While both use the same controlled syntax, only the adversarial portion incorporates sentences with structural manipulations specifically designed to be difficult for models that do not account for compositional aspects of sentence meaning. Applying this insight, we find that both the non-compositional Mean-CN model and transformer models (DefSent shown for illustration) perform similarly well on the non-adversarial portion of the new

---

7 Here and throughout the remainder of the paper we report VerbNet-CN results generated using the SemParse parser. Selected results from the GPT-4 parser are presented in subsection 8.3 in the Appendix

Table 9: Comparison of Spearman correlations of sentence similarities computed by various sentence models with human-rated sentence similarities from major STS datasets. An explanation of the models is given in Table 12 in the Appendix, and an explanation of the datasets is given in Table 5. Model types are separated by a horizontal line, from top down: vector-based arithmetic models using ConceptNet word embeddings, vector-based neural network models, syntax-based models, and hybrid models. The highest correlation for each dataset is shown in bold. STSb-capt: STSb-captions, STS3k-all: the entire STS3k dataset, STS3k-non: STS3k-non-adversarial, STS3k-adv: adversarial portion of STS3k.

Model Name	STSb-capt	STSb-test	SICK	STSS-131	STR-2022	STS3k-all	STS3k-non	STS3k-adv
Mean-CN	.806	.689	.597	.871	.612	.368	.800	-.291
Mult-CN	.260	.169	.273	.274	.057	.096	.450	-.333
Conv-CN	.164	.158	.268	.078	.057	-.042	.323	-.462
InferSent	.798	.661	.663	.868	.657	.445	.830	-.088
USE	.881	.795	.702	.900	.746	.442	.824	-.071
ERNIE-0	.619	.550	.601	.713	.592	.423	.799	-.206
ERNIE-12	.604	.549	.597	.809	.617	.576	.834	.227
SentBERT	.929	**.836**	.804	.939	.821	.580	.866	.145
DefSent	.903	.812	.785	.942	.779	.701	.868	.494
OpenAI	.923	.835	.805	**.960**	**.847**	.598	**.890**	.184
AMR-SMATCH	.565	-	.502	.653	.435	.424	.666	.029
AMR-WWLK	.738	-	.633	.829	.618	.316	.710	-.270
AMRBART	.699	.621	.637	.800	.616	.490	.837	.053
S3BERT	**.931**	.841	**.811**	.940	.826	.571	.865	.122
AMR-CN	.391	-	.517	.434	-	.602	.631	.608
VerbNet-CN	.565	-	-	-	-	.672	.652	**.647**

dataset, with correlations of 0.800 and 0.868 respectively, a difference of only 0.068. By contrast, on the adversarial portion the non-compositional Mean model performs poorly, achieving below-chance levels with a negative correlation of -0.29, while the correlation of DefSent falls to 0.49. Because of the very low performance of the non-compositional baseline model, the difference between the Mean-CN and the DefSent transformer reaches 0.8 (Figure 3). Since the adversarial and non-adversarial portions of the STS3k dataset are otherwise similar, these results demonstrate that unlike existing STS datasets, STS3k is able to discriminate between the non-compositional Mean-CN model and other models of interest. The fact the difference emerges only for the adversarial portion of the dataset indicates that the dramatic change in performance is due to the introduction of structural manipulations as discussed in subsection 4.1. This highlights the importance of utilising stimuli which can adequately probe the importance of sentence structure by controlling for the confound of lexical similarity.

Novel hybrid models outperform all other models on the STS3k dataset. We now compare the performance of the more advanced vector-based, syntax-based, and hybrid models on our new STS3k dataset. All neural network models and some syntax-based ones (e.g.
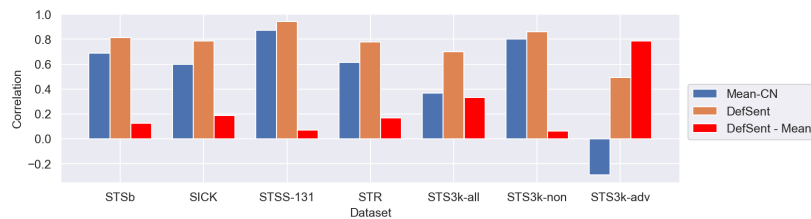
Figure 3: Comparison of correlations of a non-compositional method (the Mean-CN model) and the best performing neural network method (the DefSent transformer) on three existing STS datasets and the STS3k dataset. For the STS3k dataset, correlations are shown for the full dataset, for the non-adversarial portion, and for the adversarial portion. The difference between these two values (DefSent - Mean) provides a measure of the degree to which sentence structure (as measured by DefSent) contributes to similarity scores above and beyond lexical similarity.

S3BERT) provide very good predictions of human similarity judgments on the non-adversarial portion of the dataset (STSk-non in Table 9). However, on the adversarial part of the dataset, most transformer models show very low correlations of less than 0.2 (Table 9; see discussion below for further details). The syntax-based models also perform fairly poorly, with negative or low positive correlations. Only our two novel hybrid models, AMR-CN and VerbNet-CN, achieve similar high correlations for both subsets of the STS3k dataset (around $0.6 - 0.65$). These results highlight the superiority of the hybrid models to the other vector-based and syntax-based models in capturing human compositional representation of sentence meaning.

To further elucidate the differential performance of the various models of sentence meaning, we quantitatively compare their performances on the non-adversarial portion of the STS3k dataset (STS3k-non) to the adversarial portion only (STS3k-adv, see Figure 4). If a model has a much higher correlation with the non-adversarial dataset than with the adversarial portion, this means the model has difficulty when compositional aspects of sentence meaning become prominent. As expected, the entirely non-compositional Mean model shows the greatest difference in correlations of about 1.1. Older neural network models, including InferSent, USE, and the first layer of ERNIE, achieve somewhat lower scores of around 0.9. More recent transformer models, including SentBERT and OpenAI, along with hybrid models like S3BERT, have lower differences of around 0.7, while the best-performing transformer model (DefSent) only has a difference of 0.36. The lowest differences of all, close to zero, are shown by our novel hybrid models AMR-CN and VerbNet-CN. These results show a general trend of newer and larger neural network models exhibiting improved compositional capabilities, but the hybrid models show by far the greatest ability to incorporate compositional aspects of sentence meaning.

Transformers are insufficiently sensitive to sentence structure. To illustrate the reason for this divergence in performance, in Figure 5 we plot the human-rated sentence similarities for selected subsets of the STS3k adversarial dataset against the model cosine similarities for the Mean-CN, OpenAI, and VerbNet-CN models. The Mean-CN plot shows that for an entirely non-compositional model, only lexical similarity affects sentence similarity. As expected, this results in nearly all sentences with high lexical similarity,
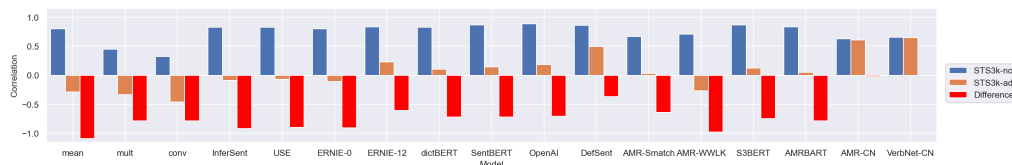
Figure 4: Comparison of correlations of all models against the non-adversarial portion of the STS3k dataset (STS3k-non) and the adversarial portion of STS3k (STS3k-adv), along with the difference between these two values. More negative values of the red bar indicate greater difficulty in modelling sentence meaning when compositional aspects are important.
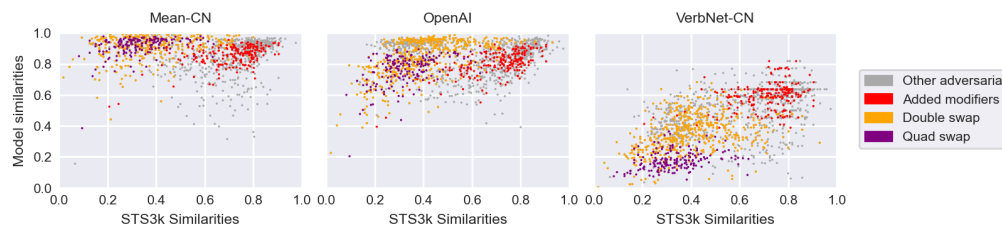

Figure 5: A comparison of three models of sentence meaning showing model cosine similarities on the vertical axis and human-rated sentence similarities on the horizontal axis. The colours highlight different subsets of the STS3k-adv dataset. Grey: all sentence similarities from the adversarial portion, purple: quads, orange: singles, red: new modifiers.

including quadruple swap, double swap, and added modifier sentence pairs, receiving high similarity scores. By contrast, the VerbNet-CN model provides similarity ratings much closer to human participants, with quadruple swaps being rated the least similar, double swaps receiving moderate similarity ratings, and added modifiers receiving highest similarity ratings. OpenAI Embeddings perform somewhat better than Mean-CN, with quadruple swaps receiving lower similarity ratings than double swaps, but overall the pattern is comparable to Mean-CN and constitutes a poor match to human ratings. These discrepancies highlight that even a sophisticated transformer model like OpenAI has not constructed sentence embeddings which reflect the core structural elements of the sentences. Swapping multiple sentence elements has little effect on cosine similarities, even though humans judge the resulting sentences to be very different in meaning. We discuss these trends across different types of sentences more systematically in subsection 5.4.

5.3 Fine-tuning against the STS3k dataset

In the previous section, we evaluated the representations of different sentence models *without* specific training on the STS3k dataset. In this section, we consider the impact of fine-tuning sentence representations against the STS3k data. Specifically, we compare the best-performing neural network transformers (SentBERT, OpenAI, ERNIE, and

DefSent) to the VerbNet-CN hybrid model, and also include as the non-compositional
Mean-CN word embeddings for comparison.

Interrogating model performance using fine-tuning. One problem with training neural net-
work models on sentence embeddings is that the embeddings of different models have
differing numbers of dimensions, and hence the resulting neural network models have
different numbers of parameters for the learned weights, which can act as a confound
(Eger, Rücklé, and Gurevych 2019). We control for this confound by using PCA to
reduce the dimensionality of model embeddings (Ferrone and Zanzotto 2020), retaining
the eight leading PCA components of each model to match the eight parameters of
VerbNet-CN. We then used this low-dimensional representation to train neural network
models with varying numbers of parameters to predict human sentence similarity
judgements in the STS3k dataset. The purpose of varying the number of parameters is to
determine the difficulty of learning the mapping between the model PCA components
and the human similarity judgements. We also note that while it is likely that more
sophisticated methods than PCA could show improved performance, our purpose here
is only to examine whether the transformers were able to learn the adversarial portion
of the STS3k dataset when trained on the non-adversarial data, not to compare differ-
ent dimensionality reduction methods[8]. We perform this analysis using two different
test/train splits. The first is simply a random split over the entire dataset. The second
uses an adversarial split, with the non-adversarial subset used for training and the
adversarial subset used for testing. This second split provides a much stronger test of the
compositional capabilities of each model by forcing out-of-distribution generalisation.

Transformers do not learn generalisable similarity ratings. The results of this fine-tuning
are shown in Figure 6. We observe that for the random split (left subfigure), the hy-
brid VerbNet-CN model shows fairly consistent correlations of around 0.7, with little
change when the number of parameters increases. By contrast, the transformer models
(SentBERT, OpenAI, DefSent, and ERNIE) show very low correlations of around 0.2
– 0.3 with few parameters, but as the number of parameters increases, the difference
in correlation narrows considerably. With enough parameters, all models can predict
human judgements with correlations of 0.7 – 0.8. By comparison, none of the trans-
formers could learn the task when trained on the non-adversarial set and tested on the
adversarial set of the STS3k dataset (right subfigure). The performance of the VerbNet-
CN model did not improve significantly with training, though it also did not degrade
and maintained at a fairly high correlation of around 0.6 regardless of the number
of parameters. These results indicate that with enough parameters and training on a
random training/testing split, all models can perform well on the testing set. However,
when there are few parameters or when trained only on the non-adversarial portion of
the STS3k dataset and tested on the adversarial portion, transformer models perform
very poorly and cannot learn the task. This constitutes evidence that, unlike humans,
the sentence representations of the transformer models we tested do not readily support
compositional generalisation to sentences different to those seen in their training set.

    We found similar results when fine-tuning a full neural network model without any
dimensionality reduction. In this case, we used SentBERT, the best-performing sentence
transformer for which fine-tuning was possible, and compared the results to the fine-

---

8 We also show in subsubsection 8.3.1 in the Appendix that similar results are observed when we perform
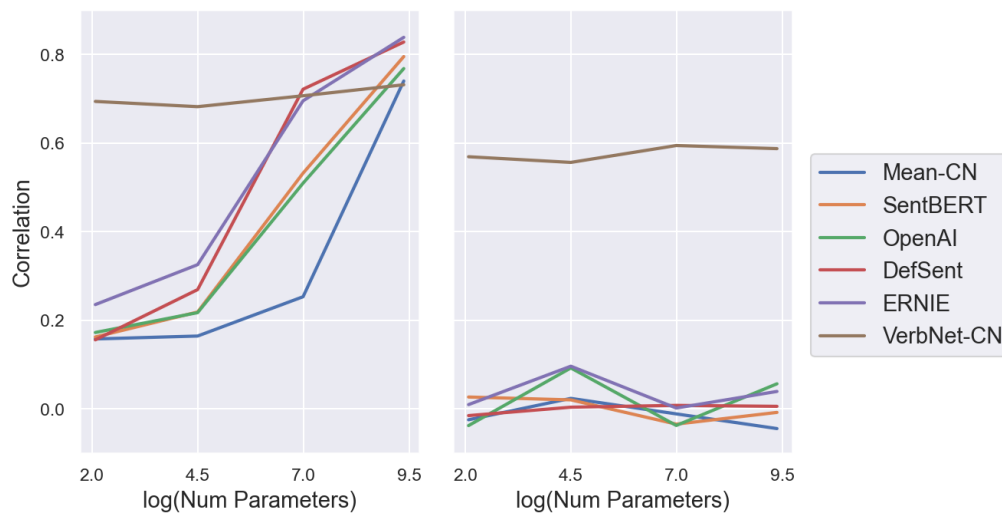 the same analysis without dimensionaity reduction

Figure 6: Correlations between model-predicted and human similarity judgements (vertical axis) against the number of parameters of the neural networks used for fine-tuning (horizontal axis). The left subplot corresponds to a random test/train split. The right subplot shows results after training on non-adversarial sentences and testing on the adversarial sentences.

tuned hybrid VerbNet-CN model (Figure 7). When neither model was given any specific training on the STS3k dataset (no training), SentBERT performed very poorly, with a correlation of only 0.17 compared to 0.65 for VerbNet-CN. When both models were fine-tuned on a representative subset of the entire STS3k dataset (random split), both achieved high correlations of around 0.8-0.85. Most interestingly, when each model was fine-tuned only on the non-adversarial portion of the dataset and evaluated on the adversarial portion (adv split), SentBERT achieved only a modest increase to 0.4, while VerbNet-CN slightly decreased to 0.57. These results indicate that even a complex transformer model trained specifically to learn sentence representations and fine-tuned on a similar dataset has limited ability to generalise to adversarial example sentences. By contrast, our hybrid VerbNet-CN model can represent the structure of such adversarial sentences even without any training.

5.4 Analysing different sentence components

We conducted additional analyses on the STS3k dataset to better understand why some models perform much better than others. We hypothesised that the best-performing models more accurately represent sentence structure, particularly how word order affects sentence meaning and the logical relation between sentence components. One way to measure this while controlling for lexical similarity is to interchange two words in a sentence (e.g. the subject and the object), thus altering sentence structure while largely preserving the constituent words. Figure 8 shows the rated similarity of sentence pairs categorised by the type of sentence manipulation, along with the predicted similarity from various compositional models. Smaller structural changes to the sentence are shown on the left, while progressively larger structural changes are shown farther to
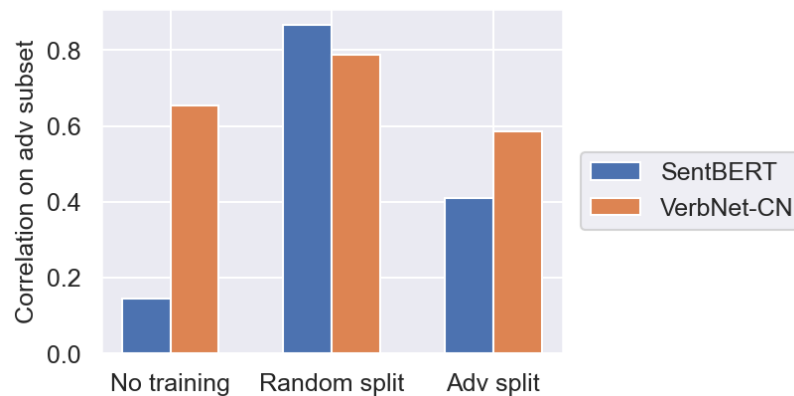
Figure 7: Correlations between STS3k-adv and the SentBERT (blue) and VerbNet-CN (green) models, with three different methods of training.

the right. Note that we opted to position negation on the far right of the graph even though it involves few structural changes, as humans are known to rate antonyms and negated sentences as highly dissimilar (Fodor, De Deyne, and Suzuki 2023). The results show that human judgments are sensitive to the number of substitutions in a monotonically decreasing fashion, while the non-compositional Mean model and transformer models (SentBERT, OpenAI, and DefSent) show relatively little sensitivity to changes in sentence structure. The VerbNet-CN hybrid model, and to a lesser extent, the DefSent transformer, show an intermediate pattern in between the other transformers and human judgements.

To quantify the difference between the models, we computed the mean absolute deviation from the normalised model similarities and the normalised human judgements across all eight categories of sentence pairs shown in Figure 8, with higher values indicating a poorer match. The entirely non-compositional Mean-CN embeddings had the highest score of 0.37, with the SentBERT and OpenAI embeddings having similarly lower scores of 0.29, 0.30 respectively. The DefSent score is lower still at 0.23, while VerbNet achieves the lowest score of 0.18, with the poor performance on Paraphrase and Negative sentence pairs partly offsetting the strong performance on Single and Swap sentence pairs. These results support our hypothesis that models that better match human similarity judgements are those with greater sensitivity to sentence structure.

Finally, to investigate whether some types of modifiers have more of an effect on sentence meaning than others, we analysed the effect of introducing a single sentence modifier on sentence similarity ratings. As shown in Figure 9, most types of modifiers have similar effects on rated sentence similarity, decreasing human-assessed similarity by an average of 0.216. The only category to show a significant difference from this was Passive with an average reduction of only 0.097, which is 2.4 standard deviations from the overall mean across categories.
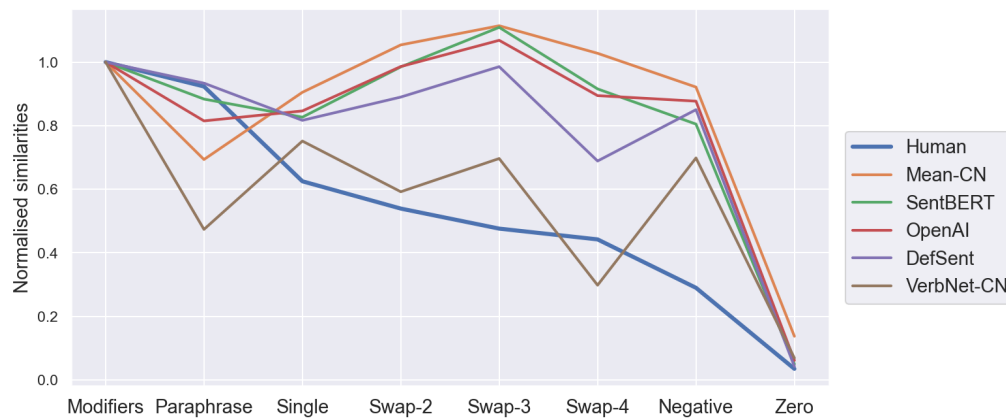
Figure 8: Mean human similarity ratings and model cosine similarities (vertical axis) plotted by the type of sentence pair in the STS3k dataset (horizontal axis). See Table 7 for an explanation of each type of sentence pair. Here we abbreviate single changes as 'single', double swap as 'swap-2', triple swap as 'swap-3', and quadruple swap as 'swap-4'. Similarities are divided by the value for 'modifiers' sentences to emphasise relative changes within each model.
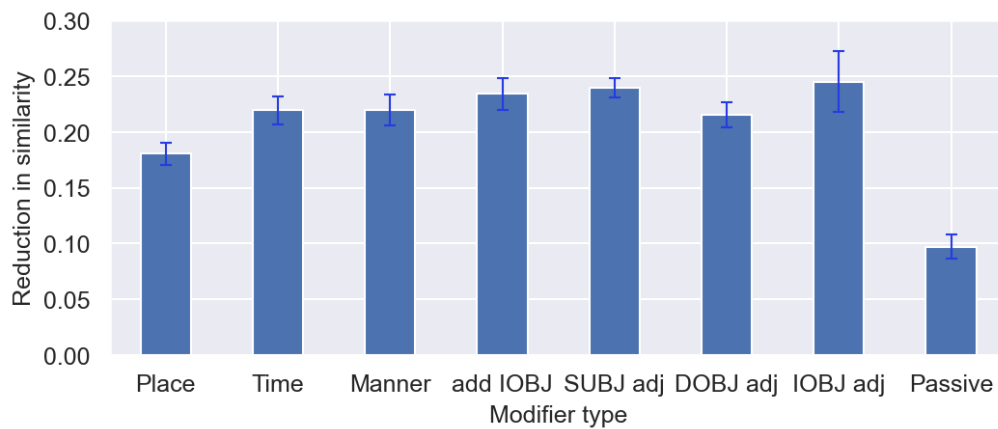

Figure 9: Change in the mean rated similarity of sentence pairs by the type of modifier added (including only sentences with a single added or changed modifier), along with the change in cosine similarities of various compositional models. Error bars denote 95% confidence intervals over sentences in each category. Add IOBJ: add an indirect object, SUBJ adj: add or change an adjective modifying the subject, DOBJ adj: add or change an adjective modifying the direct object, IOBJ adj: add or change an adjective modifying the indirect object.

6. Discussion

6.1 Summary of major findings

A major goal of computational semantics is to develop formal models to describe how humans understand and represent the meaning of words and sentences. Any such models must account for not only human comprehension of individual word meanings (lexical meaning), but also for how humans are capable of integrating familiar words in a systematic manner to understand a wide range of complex sentences they have not previously encountered (compositionality). In this study we analysed competing models of sentence meaning against human behavioural data to assess how adequately these models capture human capabilities of sentence comprehension. In particular, we investigated how well different models can capture human judgments of sentence similarity, thereby assessing the extent to which these models adequately encode sentence structure beyond the meaning of individual worlds. Since similarity is a fundamental component of any cognitive theory of representation, central to functions such as analogy, categorisation, and semantics (Goldstone and Son 2012), comparing the degree to which models of sentence meaning can capture human judgements of sentence similarity provides an important test of their adequacy as cognitive models.

To this end, our study makes four major contributions. First, we introduced a novel STS dataset (termed STS3k) constructed for the purpose of evaluating the compositional capabilities of models of sentence meaning. This dataset differs from existing STS datasets in that it contains an adversarial portion designed specifically to test whether models of sentence meaning are capable of encoding sentence structure and semantic relations beyond individual word meanings. Second, we presented a simple method for combining syntax- and vector-based semantic models into a hybrid representation that can be evaluated alongside vector-based models on STS tasks. Third, we conducted a comparative analysis of vector-based, syntax-based, and hybrid models against our novel STS3k dataset, showing that even state-of-the-art vector-based models (e.g., transformer neural networks) perform very poorly on the adversarial portion of our dataset, while our novel hybrid models succeed with no specific training. Fourth, we show through a more detailed analysis of our novel dataset that the reason why existing models perform poorly is because they are not sensitive to changes in sentence structure in the way humans are. We summarise these contributions in Figure 10.

6.2 Limitations of neural network models

Our results demonstrate that vector-based neural network models of sentences, including state-of-the-art transformers like OpenAI Embeddings, represent sentence meaning differently to human raters, which impedes their ability to perform compositional generalisation. As shown in Figure 4, the performance of these models declines dramatically when evaluated on an adversarial portion relative to a non-adversarial portion of the STS3k dataset. Because the adversarial dataset contains sentence pairs with similar lexical semantics but differing meanings by virtue of changed structure or semantic roles, and is otherwise similar to the non-adversarial portion of our dataset, this indicates the decline in performance is specifically due to the adversarial alterations, such as swapping the role of words within a sentence. This constitutes evidence that even leading transformers rely primarily on lexical cues for assessing sentence similarity, and are not very sensitive to structural changes that preserve lexical similarity while altering
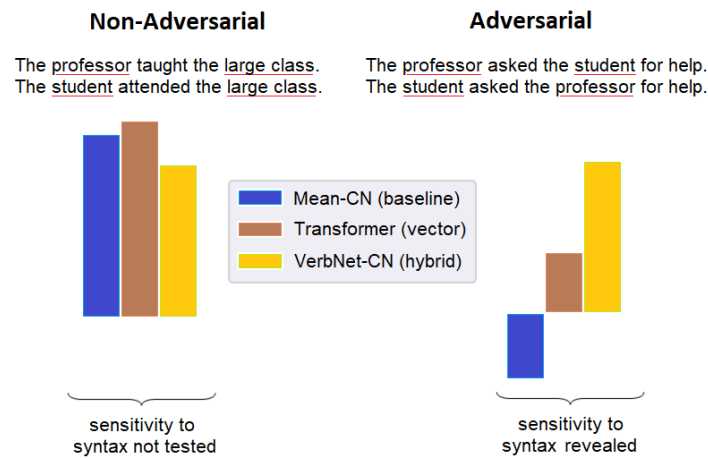
Figure 10: Schematic diagram illustrating the major contributions of our study, specifically how the contrast between the non-adversarial and adversarial portions of the STS3k dataset allows for discriminating models of sentence meaning, and illustrating how our novel VerbNet-CN hybrid model highlights how models sensitive to semantic roles can be used to understand human representation of sentence meaning.

overall sentence meaning. This suggests transformers do not adhere to principles of compositionality when producing sentence embeddings.

Since neural networks are universal function approximators (Hornik, Stinchcombe, and White 1989), we would expect that given sufficient data, they could learn to predict sentence similarity accurately for many different types of sentences. Indeed, our results in Figure 6 show that when trained on a random test/train split, transformers can learn the task well, achieving correlations of 0.7 – 0.8 with human judgements. This corroborates findings from compositional evaluation tasks such as COGS and SCAN, where standard transformer neural networks can learn the tasks fairly easily when trained on a representative range of examples but not when tested on generalisations of problems beyond that on which they were trained (Loula, Baroni, and Lake 2018; Ontanon et al. 2022; Yao and Koller 2022). A related finding is that the generalisation of transformers is often highly variable and inconsistent across training instances of the same model (McCoy, Min, and Linzen 2020), which aligns with our observation that transformer models trained only on non-adversarial sentences have difficulty generalising their performance to out-of-sample adversarial sentences.

While our findings are novel for STS tasks, several previous studies have found analogous results of limited compositional capability when controlling for lexical overlap using paraphrase data (Yu and Ettinger 2020; Bernardi et al. 2013). Other investigations have found that transformer neural network models perform highly inconsistently on subtle variations of language tasks that humans would regard as equivalent (Srivastava et al. 2022; Dankers, Bruni, and Hupkes 2022), indicating that they have not learned to perform the task in a manner comparable to humans. Whether this limitation could be overcome with a much larger training dataset of sentences covering a wider range of topics and sentence structures is unclear. Based on previous work, it is likely that transformers will struggle to generalise to sentences significantly different from

those in the training distribution, and given that language is necessarily productive in generating sentences of arbitrary length and combinations, presenting a wide enough range of sentences may be infeasible. This highlights the importance of adversarial testing to investigate whether models extract the relevant features that will enable them to perform language tasks across various contexts.

The only neural network model to show moderate correlations on the adversarial portion of the dataset is the transformer DefSent, which achieves a surprisingly high correlation of 0.49 despite performing at or slightly below the level of SentBERT on the other datasets. Judging from Figure 8, this is due to DefSent giving lower similarities for single, double, triple, and quadruple sentence pairs relative to modified sentences, which is closer to human judgments than any of the other transformer models. We speculate that the superior performance of DefSent may be due to its unique training, in which it learns to map a word to its definition sentence from a lexical dictionary. However it is unclear exactly why this training method would lead to such an improvement in performance on the adversarial task.

Another novel result from our analysis is that transformers, at least of the scale assessed in this study, do not efficiently extract semantic information from word order. Much of the adversarial aspect of our STS3k dataset relies on varying the position of words within a sentence. For example, in one version, we move a word from the subject position near the start of the sentence to the object position near the end of the sentence. As we show in Figure 8, such sentence alterations have little effect on the similarities computed from transformer embeddings, indicating that transformers are not sensitive to such changes. Since transformers use positional encoding to represent the linear ordering of words within a sentence, semantic role information should be readily available to the transformers (Dufter, Schmitt, and Schütze 2022). However, our results indicate that the transformers in our benchmark have difficulty extracting this information from positional embeddings. We speculate that this may result from transformers relying on lexical information and other incidental correlations for next-word or masked token prediction tasks, meaning that the underlying structural and semantic role information from the sentence is underutilised. Although numerous probing studies have found that transformers do represent information about syntax and word order in their hidden layers (Clark et al. 2019; Manning et al. 2020), this information may not be effectively utilised in sentence embeddings for representing the meaning of the entire sentence.

The fact that the transformers investigated in this study fail to match human predictions of sentence similarity does not mean that transformers are useless as language models. The transformer architecture is still very flexible and underpins many models which are highly successful in numerous language tasks. Rather, our results are significant because they show that, regardless of their success on downstream language tasks, transformers (along with other vector-based models) are insufficiently sensitive to sentence structure, and hence do not represent sentence meaning in the manner that humans do. As we show in Figure 5 and Figure 8, transformer sentence embeddings do not vary in proportion to the degree of structural change within a sentence (e.g. when words interchange their semantic roles). This indicates that they have fundamentally failed at the task of representing sentence meaning in a manner which respects well-established psychological and linguistic principles relating to the effects of sentence structure on meaning.

Our findings align with the results of various recent studies demonstrating the limitations of transformers as plausible models of human compositional language processing. Gupta, Kvernadze, and Srikumar (2021) found that performing various transformations on input sentences, such as randomly shuffling the word order, resulted

34

in only small changes to the predictions made by BERT-family transformers on a range
of NLI tasks, despite the fact that the resulting sentences were now entirely meaningless.
Golan et al. (2023) constructed a set of 'controversial' sentence pairs for which different
models disagreed about which sentence of the two was most likely. They found in
a series of tests that all transformers displayed behaviour inconsistent with human
judgements. Webson and Pavlick (2022) found that even for very large transformer
models like GPT-3, there was little to no difference in performance on various NLI tasks
when instructive prompts were used compared to nonsensical or irrelevant prompts,
casting doubt on whether models are capable of understanding such prompts in a
human-like manner. Another study found a similar result using negated prompts (Jang,
Ye, and Seo 2023). Various other techniques involving injecting irrelevant content into
prompts or modifying prompts in ways which do not change their meaning (such as
simple typographical substitutions) have likewise highlighted that transformers do not
appear to understand the meaning of their prompts (Jiang, Chen, and Tang 2023; Wang
et al. 2023; Shi et al. 2023). These difficulties likely result from the fact that transformers
primarily rely on superficial heuristics and spurious correlations learned from their
training data, allowing them to perform well on many typical tasks even without
forming relevant structured representations of the situation or problem to be solved
(Niven and Kao 2019; Zhang et al. 2022; Dziri et al. 2023; Gubelmann and Handschuh
2022b). Our results provide further evidence in support of this general conclusion, high-
lighting that transformers do not form structural representations of sentence meaning
capable of capturing the sorts of information important to human representations of
sentence meaning. This limits the value of transformers both as psychological models
of representations of sentence meaning, and also on tasks requiring extensive capability
with generalisation or compositional reasoning.

6.3 Integrating vector-based and syntax-based methods

Our novel hybrid models differ in important ways from previous methods of combining
vector-based and syntax-based models. Most traditional hybrid models attempt to inject
syntax into neural networks by training them to perform graph prediction tasks. As out-
lined in Table 4, this approach has typically led to only modest increases in correlation
with human data, though it is unclear if this is due to a limitation of the methodology
or existing STS datasets. Furthermore, such approaches have been criticised as theoreti-
cally unmotivated, as there is typically little explanation of what the embedding space is
intended to represent. One study has suggested sentence embeddings share a semantic
space with individual words, with both the direction and length of sentence embed-
dings conveying semantic information (Amigó et al. 2022). However considering that
embeddings in transformer neural network models are known to be highly anisotropic,
meaning that a few dimensions account for nearly all of the vector length (Timkey
and van Schijndel 2021; Su et al. 2021), it seems unlikely that embeddings learned by
transformers represent sentences in this way. Our approach differs in not representing
a sentence using a single vector embedding, but instead utilising a hybrid method in
which individual words are represented using static word embeddings, which are then
combined in a manner specified by a syntax-based model to form the full sentence
representation. The result is not a single embedding for the entire sentence but instead
a structured representation, the elements of which consist of embeddings of sentence
components.

   Our VerbNet-CN and AMR-CN hybrid methods are not consistently superior to
transformers, as they perform worse both on existing STS datasets and the STS3k non-

adversarial portion (Table 9). This is unsurprising, given that each semantic role uses simple averaged word embeddings rather than the sophisticated attention mechanism of transformers. Furthermore, our hybrid models are much less flexible than transformers, designed only to extract a defined set of semantic roles in relatively simple single-clause sentences. Many aspects of natural language, including auxiliary verbs, multiple clauses, polysemy, and multi-word expressions, are not incorporated. As such, the purpose of our novel hybrid models is not to replace transformers or even achieve comparable performance on downstream language tasks, but rather to highlight the inadequacy of current transformer models in representing sentence structure, and to illustrate the value of explicitly representing elements of sentence structure such as semantic roles. Our aim is for these models to serve as a simple baseline method for more complex models which vector-based and syntax models incorporating a wider range of syntactic and semantic components.

One question raised by our analysis is the status of semantic roles or predicate arguments in the context of vector-based models of semantics. How exactly are they to be interpreted? One possibility is that semantic roles correspond to high-level semantic features, which together characterise the semantic meaning of the sentence. However, a problem with this interpretation is that features in semantic space are typically represented as independent dimensions which can vary separately from one other. By contrast, semantic roles or arguments of predicates are "roles" that bind to their "fillers" in each particular context. There has been extensive discussion about how to integrate symbolic role-filler dynamics with vector-based representations (Soulos et al. 2019; Vankov and Bowers 2020), with tensor products being a recent popular approach (Badreddine et al. 2022; Smolensky et al. 2022). We leave this aspect to future research.

A second question raised by our analysis is how it can be established what the "correct" sentence representation structure is. Which semantic roles are the most important in describing human semantic representations? In this study, we adopted a heuristic approach of selecting major semantic roles based on the VerbNet framework, as discussed in subsection 4.6. However, we do not make any claim that the eight we have selected are the singular "correct" semantic roles. Indeed, it is likely that different roles are important in different contexts and domains, though some are likely prominent and broadly applicable. Our findings do not suggest that any specific semantic roles are psychologically real. Instead, we claim only that incorporating semantic roles into sentence representations improves their fit to human judgements, and this constitutes evidence that such structured elements of meaning form part of human representations of sentence meaning.

Given these considerations, we affirm previous calls for the importance of combining syntax-based and vector-based approaches to language modelling, ensuring that vector-based models are equipped with the appropriate inductive biases to facilitate learning representations and identifying features that will be useful beyond the training set.

6.4 Human representation of sentence meaning

Research into compositional semantics is hampered by a lack of agreement on how compositionality should be characterised (Pagin and Westerståhl 2010; Szabó 2012). Indeed, it has been argued that the concept is formally vacuous without being tied to a particular syntactic formalism or set of rules (Janssen 1986; Zadrozny 1994; Westerståhl 1998). Furthermore, human language is unlikely to adhere to the strict rules of compositionality. If it did, different words fulfilling the same abstract role should be processed in exactly

the same way irrespective of the lexical meaning of the word, whereas in fact contextual and situational effects have a significant effect on how humans represent and process sentence meaning. As such, it may be helpful to think of compositionality as a rough abstraction describing some idealised aspects of cognitive and linguistic competencies rather than as a strict formal definition (Dankers, Bruni, and Hupkes 2022; Martin and Baggio 2020). We adopt this heuristic approach to analysing competing models of sentence meaning, asking whether such models construct sentence representations that facilitate inferences and behaviours characteristic of compositional systems, such as generalisation and systematicity.

Our results show that raters are sensitive to subtle and non-obvious distinctions between sentences, and make discriminating decisions even in this vaguely defined task. This is most evident from Figure 8, where humans give similarity ratings of 0.6 for two sentences with a single element altered, decreasing progressively to 0.4 for sentence pairs with four elements interchanged. This aligns with previous results supporting an 'edit-distance' approach for assessing sentence similarity, whereby humans judge sentence similarity based on the number of sentence elements (such as semantic roles) that are altered between the two sentences (Gershman and Tenenbaum 2015; Kemp, Bernstein, and Tenenbaum 2005). In Figure 9 in the section 8, we show that participants were roughly equally responsive to additions of all different types of modifier elements, each of which reduced assessed sentence similarity by about 0.2, except for the use of passive voice which only reduced similarity by about 0.1. This latter result is especially interesting since, in terms of truth conditions or logical entailments, sentences expressed in the active and passive voice are identical, the only difference being emphasis and connotation. The fact that humans assess such sentence pairs as differing in meaning highlights the limitation of representational approaches that ignore such subtle but important aspects of meaning. We also note that human similarity judgements are sub-additive in the number of modifiers included (see Figure 9), with each modifier having a larger effect on similarity when occurring individually than when combined with others.

Previous work in linguistics and cognitive psychology has demonstrated that humans are sensitive to the roles played by words within a sentence (Philipp et al. 2017; Lau, Clark, and Lappin 2017; Alishahi and Stevenson 2010). However, it is unclear exactly how such roles and structure are represented or encoded, whether this takes the form of a static set of roles that are widely reused across contexts (such as 'agent' or 'patient') or a set of selection rules linking different verbs with their common arguments. Another approach models human representations as frames, which are highly structured representations evoked by an entire situation. While our results do not allow us to distinguish between such models, insofar as a model utilising a simple set of semantic roles shows a much higher correlation with experimental data than models that do not, there is evidence that some type of semantic role or structure plays a role in human judgments of sentence meaning.

Experimental results have shown that humans can readily learn novel categories and predicates with only a few examples by using various inductive biases (Lake, Linzen, and Baroni 2019). However, current methods for training neural networks do not typically incorporate such inductive biases either directly through architectural constraints or indirectly in the way they are trained. As such, while they form representations suitable for word prediction and which generalise to other inference tasks, these representations are typically unsuitable for tasks involving substantial generalisation or systematic variation of components beyond the training data.

7. Conclusion

In this paper, we introduced a novel semantic textual similarity dataset involving adversarial sentence pairs designed to test for compositional representations of sentence meaning while controlling for lexical similarity. We then tested various models against this dataset, including vector-based, syntax-based, and hybrid models. We found that for the adversarial subset of our task, existing vector-based and syntax-based models failed to accurately predict human judgements of semantic similarity, while our novel hybrid model performs well. Our analysis of these results has shown that while humans rate sentence similarity in accordance with the semantic roles of different sentence components, existing vector-based models, including state-of-the-art transformer neural network models, do not represent sentence structure in this way and perform poorly on the adversarial portion of the dataset. The transformers could only learn the task when trained on adversarial examples, but could not generalise from the non-adversarial to the adversarial portion of the dataset. We further showed how syntax-based approaches to sentence representation can be combined with vector-based static word embeddings to produce a hybrid method that performs substantially better than any transformer model on the adversarial dataset. Overall, our findings highlight the limitations of existing transformer models of sentence representation, and the value of semantic roles and structural information in describing human representations of sentence meaning.

8. Appendix

8.1 Supplementary background
8.1.1 Further explanation of vector/syntax/hybrid terminology. Unfortunately there is no standard terminology or categorisation for describing different approaches to modelling sentence meaning. In this paper we attempt to simplify our presentation by focusing on two broad classes of models, which we term *vector-based* and *syntax-based* semantics. A third class, which attempts to integrate aspects from both approaches to combine their respective strengths, we term *hybrid approaches*. We adapted these terms from Žabokrtský, Zeman, and Ševčíková (2020), who distinguish between 'deep-syntactic' and 'vector space' models of sentence meaning. We intend these labels to roughly separate differing approaches to representing sentence meaning in a manner that simplifies and provides structure to the presentation of our results, whilst also acknowledging alternative classification terminology. For example, in their insightful review, Liang and Potts (2015) use the terms 'distributional representations' and 'semantic parsing', while Ferrone and Zanzotto (2020) distinguish between 'distributed' and 'symbolic' sentence representations. We do not intend our terminology to provide an exhaustive or strictly dichotomous categorisation of all models of sentence meaning.

Syntax-based and vector-based models are typically evaluated differently. In particular, syntax-based models are usually evaluated by comparing the sentence representations with a gold standard of human-annotated sentence parses. By contrast, vector-based models are assessed using a range of tasks including natural language inference, paraphrase, translation, sentiment analysis, and semantic similarity tasks. This difference in evaluation methods stems from slightly different objectives and strengths of different types of models (Beltagy et al. 2016; Ferrone and Zanzotto 2020). Syntax-based methods usually focus on producing a graph-based parse of a sentence, and require augmentation to perform text generation or other forms of NLI inference. By contrast, most vector-based models do not intrinsically have any representation of syntax which

can be compared to a human-annotated sentence parse, and are instead trained directly to perform next word prediction or some other linguistic task.

8.1.2 Other textual similarity datasets. Beyond sentence similarity ratings, several other datasets exist pertaining to related tasks, including similarity judgments of adjective-noun bigrams (Vecchi et al. 2017; Asaadi, Mohammad, and Kiritchenko 2019; Cordeiro et al. 2019), sets of sentence paraphrases (Dolan and Brockett 2005), or pairs of sentences differing by grammatical acceptability (Warstadt et al. 2020). In this paper we do not analyse these data, restricting our scope to datasets containing similarity or relatedness judgements of full sentences. Bigram similarity captures only a small part of sentence meaning, while sentence paraphrase data only explores one extreme on the range of similarity, and so likewise is of limited use for our purposes. Grammatical acceptability datasets primarily probe the ability of language models in various grammatical domains, such as determiner agreement, verb conjugation, and quantifiers, which are also of less direct relevance to assessments of sentence meaning than STS datasets.

8.1.3 Criticisms of Semantic Textual Similarity tasks. STS has been criticised for not being reliably predictive of model performance on applied language tasks (Wang et al. 2019; Wang, Kuo, and Li 2022; Abe et al. 2022), and for being subject to other limitations such as low inter-annotator agreement (Batchkarov et al. 2016). While acknowledging these concerns, we believe it is an appropriate metric for our current study for several reasons. First, STS is one of only a few methods capable of directly comparing the internal representations of models of sentence meaning. This is of particular interest owing to recent studies highlighting that despite impressive performance of neural network models on various language tasks, the models often fail to learn or utilise generalisable representations of the underlying structure of the problem or domain in question. Instead, often the models achieve high levels of performance by extracting complex statistical artifacts and utilising heuristics which do not generalise beyond the specific dataset used for training or assessment (Gupta, Kvernadze, and Srikumar 2021; Gubelmann and Handschuh 2022b; Zhang et al. 2022). We wish to probe the internal representations of different approaches to sentence meaning to investigate how well they are able to incorporate key aspects of sentence structure. Second, it has been noted that one reason for the relatively low correlation between performance on semantic similarity tasks and performance on other downstream applications is because for many tasks (e.g. sentiment analysis or co-reference identification), only certain aspects or features of the sentence are relevant (Wang, Kuo, and Li 2022). As a holistic measure of the similarity of meaning of two sentences, STS datasets will not always correlate with performance on such tasks. Lack of overlap of vocabulary and subject domain has also been identified as a factor contributing to low predictivity (Abe et al. 2022). These issues are of less relevance since our focus is on the empirical adequacy of the representations themselves, rather than their utility for any particular downstream application. Third, as we show in subsection 5.1, results from our STS3k dataset show very high inter-rater reliability.

8.1.4 Compositional inference tasks. Compositional inference tasks are designed to test whether language models are capable of appropriately identifying structural similarities between superficially disparate inputs, and utilising this information to perform tasks that require generalisation beyond a training set. Here we summarise three major datasets in this tradition. The SCAN dataset (Lake and Baroni 2018) consists of a set of navigation commands presented in a simple English sentence, each paired with a

corresponding sequence of movement instructions. The dataset is arranged into different train-test splits, which requires compositional reasoning to construct movement instructions corresponding to a novel input sentence. The COGS dataset (Kim and Linzen 2020) consists of a series of natural language sentences randomly generated in accordance with certain structural parameters, each paired with a corresponding logical form. The objective of the task is to predict the logical form of a novel sentence. The dataset is designed so that compositional generalisation is required between the training and test sets, such as varying the grammatical role of a word or deeper recursion. Finally, the CFQ dataset (Keysers et al. 2019) consists of a series of natural language questions and the corresponding syntax for querying a structured database. The goal of the task is to construct a structured database query from a novel sentence.

Numerous studies have found that syntactic parsing models solve these compositional tasks easily, while even state-of-the-art neural network models struggle, especially for instances requiring extensive compositional generalisation (Yao and Koller 2022). Nevertheless, various strategies have been developed for modifying transformer architectures to improve compositional performance. This includes using relative position encodings (Ontanon et al. 2022), modifying the training data (Patel et al. 2022), and using longer training periods (Csordás, Irie, and Schmidhuber 2021). Most recently, it has been shown that careful choice of prompts can substantially improve LLM performance on compositional tasks (Zhou et al. 2022). Some have even argued that these techniques show, in contrast to conventional wisdom, that transformers with the appropriate training are capable of compositional reasoning (Csordás, Irie, and Schmidhuber 2021).

While compositional tasks are valuable for assessing how LLMs combine word meanings, they are nonetheless subject to several limitations. First, they are insufficiently discriminative, being simultaneously too easy for symbolic methods and too difficult for most vector-based methods. An ideal method of evaluation should discriminate the performance of both types of models, thereby enabling a more precise interrogation of their strengths and weaknesses. Second, existing tasks (involving constructing dataset queries or abstract movement instructions) are somewhat artificial and removed from human natural language performance (Dankers, Bruni, and Hupkes 2022). As such, while these tasks are suitable for testing compositionality in the abstract, they are not suited to testing competing representations of natural language sentences. Partly in response to such limitations, researchers have emphasised the importance of assessing LLMs on non-synthetic data (Dankers, Bruni, and Hupkes 2022; Yao and Koller 2022; Ribeiro et al. 2020b), with several studies showing that performance on synthetic data with highly controlled vocabulary is not always predictive of performance on less constrained, more natural tasks (Shaw et al. 2021).

8.2 Supplementary methods

8.2.1 Instructions to participants. The text below was provided to participants prior to making sentence similarity judgements.

> Please read the following instructions carefully before proceeding.
>   In this questionnaire you will be presented with a series of paired sentences. Your task is to judge how similar is the meaning of the two sentences. You will make this judgement by choosing a rating from 1 (very dissimilar) to 7 (very similar). In providing your rating, consider both the similarity in meaning of the individual words contained in the sentences, as well as the similarity of the overall idea or meaning expressed by the sentences.

Some of the sentences may be slightly unusual or ambiguous; nevertheless you should do your best to understand their likely meaning. Bear in mind that we are not looking for any one specific 'right answer' or strategy in your responses. Your task is simply to make a judgement about how similar you think is the meaning of the two paired sentences. The only exception is that if you find a sentence that truly does not make any sense at all, then you should give it a very low similarity to whatever it is paired with. In all other cases, make your best judgement based on your assessment of overall meaning of the sentences.

There is no time limit to this task, however each sentence pair should not take more than a few seconds to judge. There is no need to spend a long time pondering each sentence. In total the task should take around 20-30 minutes.

Thanks very much for your time!

8.2.2 Parsing instructions for GPT-4. The instruction below was provided to the OpenAI client using GPT4 for parsing sentences one pair at a time.

Two sentences are given below. First, identify the main verb in each sentence. Each sentence should only have a single main verb. Use simple present conjugation. Second, label the semantic roles in each of these new sentences. Use the roles: 'Agent', 'Patient', 'Theme', 'Time', 'Manner', 'Location', 'Trajectory'. Print all results in a single list on one line. Print each role regardless of whether it is found in the sentence. Do not explain your answers. Here is one example of what to print:

'Food is what people and animals reluctantly eat on Thursdays.'

{'Verb': 'is', 'Agent': 'food', 'Patient': 'NONE', 'Theme', 'what people and animals eat', 'Time': 'on Thursdays', 'Manner': 'reluctantly', 'Location': "NONE", 'Trajectory': "NONE"}

Here are the two sentences for you to parse:

8.3 Supplementary results

8.3.1 Fine-tuning without dimensionality reduction. As an additional check, we performed the fine-tuning as described in subsection 5.3, but without the dimensionality reduction. This meant that the number of parameters in each trained model is slightly different (see Table 10), as some transformers have larger embeddings than others. Qualitatively the results are similar to those shown in Figure 6, with transformer models learning the task well when trained on a random test/train split, but unable to learn the task when trained on the non-adversarial subset and required to generalise out of sample. By contrast, the VerbNet-CN hybrid model achieves moderately good performance in both versions of the task, and is relatively unaffected by the number of parameters.

Table 10: Number of parameters for each fine-tuned model, equal to the entry for each cell multiplied by the corresponding column header.

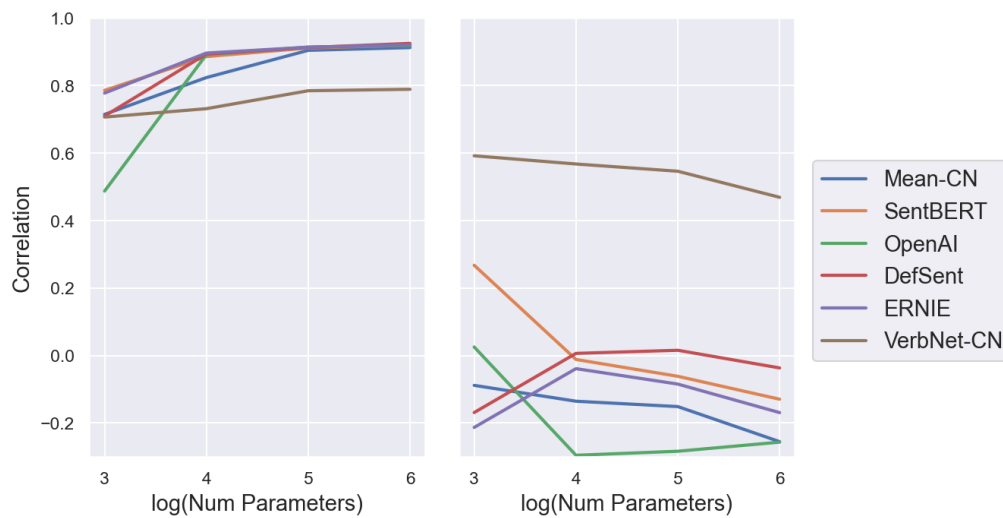| Model name	10^3	10^4	10^5	10^6
Mean-CN	.598	.599	.708	.709
SentBERT	1.536	1.537	1.647	1.647
OpenAI	2.800	2.801	2.911	2.911
DefSent	2.048	2.049	2.159	2.159
ERNIE	1.536	1.537	1.647	1.647
VertNet-CN	1.090	1.181	1.190	1.109

Figure 11: Correlations between model-predicted and human similarity judgements (vertical axis) against the approximate number of parameters of the neural networks used for fine-tuning (horizontal axis). The left subplot corresponds to a random test/train split. The right subplot shows results after training on non-adversarial sentences and testing on the adversarial sentences.

8.3.2 VerbNet-CN with GPT-4 parser. In Table 11 we show a comparison of the correlation between STS3k and the VerbNet-CN hybrid model using both the original SemParse parser, and the alternative GPT-4 parser. The correlation between the two similarity series was computed to be 0.92. These results indicate that our novel hybrid approach is robust to the particular parsing method used.

Table 11: Summary of models of sentence meaning analysed in this study.

Model name	STS-all	STS-non	STS-adv
VerbNet-CN (SemParse parser)	.672	.652	.647
VerbNet-CN (GPT-4 parser)	.673	.685	.627

8.4 Supplementary tables

The table below summarises all the models examined in the present paper. The models are grouped by category separated by horizontal lines, from top to bottom: arithmetic vector-based, neural network vector-based, syntax-based, and hybrid.

Table 12: Summary of models of sentence meaning analysed in this study.

Model name	Type	Explanation	Citation
Mean-CN	Arithmetic vector-based	Average of ConceptNet token-wise embeddings after pre-processing of sentences to remove non-content words.	Mitchell and Lapata (2010)
Mult-CN	Arithmetic vector-based	Elementwise multiplication of ConceptNet embeddings after pre-processing.	Mitchell and Lapata (2010)
Conv-CN	Arithmetic vector-based	Convolution of ConceptNet token-wise embeddings after pre-processing.	Blouw et al. (2016)
InferSent	Vector-based	A bi-directional LSTM trained on a variety of natural language inference tasks.	Conneau et al. (2017)
Universal	Vector-based	A standard transformer architecture trained on a range of language tasks.	Cer et al. (2018)
ERNIE 2.0	Vector-based	A transformer based on the BERT architecture trained using multi-task learning.	Sun et al. (2020)
SentBERT	Vector-based	The mpnet-base transformer model with additional training to predict paired sentences from a large dataset.	Reimers and Gurevych (2019)
DefSent	Vector-based	The RoBERTa-large transformer model fine-tuned using about 100,000 words paired with their dictionary definitions. We use the cls output.	Tsukagoshi, Sasano, and Takeda (2021)
OpenAI Embeddings	Vector-based	Embeddings provided from the OpenAI API, based on a large transformer with additional fine-tuning from human feedback.	Ouyang et al. (2022)
AMR-SMATCH	Syntax-based	Sentences parsed using an AMR parser, and similarity between the resulting graphs computed using SMATCH.	Cai and Knight (2013)
AMR-WWLK	Syntax-based	Sentences parsed using an AMR parser, and similarity between the resulting graphs computed using WWLK.	Opitz, Daza, and Frank (2021b)
AMRBART	Hybrid	A transformer architecture trained to encode AMR graphs.	Bai, Chen, and Zhang (2022)
S3BERT	Hybrid	A transformer based on SentBERT with extra training to use AMR graph-based metrics to construct an overall similarity score.	Opitz and Frank (2022b)
AMR-CN	Hybrid	An AMR parser produces a graph, then similarity is computed by averaging ConceptNet word embeddings for graph components.	Introduced in this paper
Verbnet-CN	Hybrid	Sentence parsed into VerbNet semantic roles, then similarity computed as average of ConceptNet word embeddings over roles.	Introduced in this paper

References

Abdalla, Mohamed, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796.

Abe, Kaori, Sho Yokoi, Tomoyuki Kajiwara, and Kentaro Inui. 2022. Why is sentence similarity benchmark not predictive of application-oriented task performance? In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 70–87.

Abend, Omri and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation*, pages 497–511, ACL (Association for Computational Linguistics).

Alishahi, Afra and Suzanne Stevenson. 2010. A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, 25(1):50–93.

Almeida, Felipe and Geraldo Xexéo. 2019. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

Amigó, Enrique, Alejandro Ariza-Casabona, Victor Fresno, and M Antònia Martí. 2022. Information theory–based compositional distributional semantics. *Computational Linguistics*, 48(4):907–948.

Asaadi, Shima, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516.

Badreddine, Samy, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. 2022. Logic tensor networks. *Artificial Intelligence*, 303:103649.

Bai, Jiangang, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-bert: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020.

Bai, Xuefeng, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for amr parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015.

Bakarov, Amir. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Baker, Collin F, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Ballesteros, Miguel, Bernd Bohnet, Simon Mille, and Leo Wanner. 2014. Deep-syntactic parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1402–1413.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Baroni, Marco. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in language technology*, 9:241–346.

Batchkarov, Miroslav, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.

Beltagy, Islam, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J Mooney. 2016. Representing meaning with a combination of logical and distributional

models. *Computational Linguistics*, 42(4):763–808.

Bernardi, Raffaella, Yao-Zhong Zhang, Marco Baroni, et al. 2013. Sentence paraphrase detection: When determiners and word order make the difference. In *Proceedings of the IWCS 2013 Workshop Towards a Formal Distributional Semantics*, pages 21–29.

Bevilacqua, Michele, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

Blacoe, William and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 546–556.

Blouw, Peter, Eugene Solodkin, Paul Thagard, and Chris Eliasmith. 2016. Concepts as semantic pointers: A framework and computational model. *Cognitive science*, 40(5):1128–1162.

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Boleda, Gemma and Aurélie Herbelot. 2016. Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619–635.

Bölücü, Necva, Burcu Can, and Harun Artuner. 2023. A siamese neural network for learning semantically-informed sentence embeddings. *Expert Systems with Applications*, 214:119103.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Cai, Shu and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Cai, Xingyu, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Chandrasekaran, Dhivya and Vijay Mago. 2021. Comparative analysis of word embeddings in assessing semantic similarity of complex sentences. *IEEE Access*, 9:166395–166408.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Chersoni, Emmanuele, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and C-R Huang. 2019. A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, 25(4):483–502.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Clark, Stephen. 2015. Vector space models of lexical meaning. *The Handbook of Contemporary semantic theory*, pages 493–522.

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Association for Computational Linguistics, Copenhagen, Denmark.

Cordeiro, Silvio, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Csordás, Róbert, Kazuki Irie, and Juergen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634.

Dankers, Verna, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Dolan, Bill and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Donatelli, Lucia and Alexander Koller. 2023. Compositionality in computational linguistics. *Annual Review of Linguistics*, 9.

Dufter, Philipp, Martin Schmitt, and Hinrich Schütze. 2022. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Eger, Steffen, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the evaluation of sentence embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60.

Emerson, Guy. 2020. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453.

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

Erk, Katrin. 2016. What do you know about an alligator when you know the company it keeps? *Semantics & Pragmatics*, 9(17):1–63.

Erk, Katrin. 2022. The probabilistic turn in semantics and pragmatics. *Annual Review of Linguistics*, 8:101–121.

Etcheverry, Mathias and Dina Wonsever. 2019. Unraveling antonym's word vectors through a siamese-like network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3297–3307.

Farouk, Mamdouh. 2020. Measuring text similarity based on structure and word embedding. *Cognitive Systems Research*, 63:1–10.

Ferrone, Lorenzo and Fabio Massimo Zanzotto. 2020. Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. *Frontiers in Robotics and AI*, 6:153.

Fodor, James, Simon De Deyne, and Shinsuke Suzuki. 2023. The importance of context in the evaluation of word embeddings: The effects of antonymy and polysemy. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 155–172, Association for Computational Linguistics, Nancy, France.

Fodor, Jerry and Brian P McLaughlin. 1990. Connectionism and the problem of systematicity: Why smolensky's solution doesn't work. *Cognition*, 35(2):183–204.

Fodor, Jerry A and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Frankland, Steven M and Joshua D Greene. 2020. Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, 71:273–303.

Frege, Gottlob et al. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

Gajewski, Jon. 2015. Foundations of formal semantics.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gershman, Samuel and Joshua B Tenenbaum. 2015. Phrase similarity in humans and machines. In *CogSci*, Citeseer.

Golan, Tal, Matthew Siegelman, Nikolaus Kriegeskorte, and Christopher Baldassano. 2023. Testing the limits of natural

language models for predicting human language judgements. *Nature Machine Intelligence*, 5(9):952–964.

Goldstone, Robert L. and Ji Yun Son. 2012. 155 Similarity. In *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press.

Goyal, Palash and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

Graves, Alex. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Grefenstette, Edward and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404.

Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. 2020. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*.

Gubelmann, Reto and Siegfried Handschuh. 2022a. Uncovering more shallow heuristics: Probing the natural language inference capacities of transformer-based pre-trained language models using syllogistic patterns. *arXiv preprint arXiv:2201.07614*.

Gubelmann, Reto and Siegfried Handschuh. 2022b. Uncovering more shallow heuristics: Probing the natural language inference capacities of transformer-based pre-trained language models using syllogistic patterns. *arXiv preprint arXiv:2201.07614*.

Gung, James. 2020. Semparse verbnet parser.

Gupta, Ashim, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert & family eat word salad: Experiments with text understanding. *Proceedings of the AAAI conference on artificial intelligence*, 35(14):12946–12954.

Hampton, James A. 2017. *Compositionality and concepts in linguistics and psychology*. Springer International Publishing.

Hartung, Matthias, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 54–64.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Jang, Joel, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62, PMLR.

Janssen, TMV. 1986. Foundations and applications of montague grammar: 8m part 1: philosophy, framework, computer science.

Jiang, Shuyu, Xingshu Chen, and Rui Tang. 2023. Prompt packer: Deceiving llms through compositional instruction with hidden attacks. *arXiv preprint arXiv:2310.10077*.

Kabbach, Alexandre, Corentin Ribeyre, and Aurélie Herbelot. 2018. Butterfly effects in frame semantic parsing: impact of data processing on model ranking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3158–3169, Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Kartsaklis, Dimitri, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 114–123.

Kasami, Tadao. 1966. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*.

Kemp, Charles, Aaron Bernstein, and Joshua B Tenenbaum. 2005. A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society*, pages 1132–1137, Citeseer.

Keysers, Daniel, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Kim, Najoung and Tal Linzen. 2020. COGS: A compositional generalization challenge

based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Association for Computational Linguistics, Online.

Kingsbury, Paul R and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

Krasnowska-Kieraś, Katarzyna and Alina Wróblewska. 2019. Empirical linguistic study of sentence embeddings. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5729–5739.

Lake, Brenden and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882, PMLR.

Lake, Brenden M, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. In *41st Annual Meeting of the Cognitive Science Society: Creativity+ Cognition+ Computation, CogSci 2019*, pages 611–617, The Cognitive Science Society.

Landauer, Thomas K, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

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

Leung, Wai Ching, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115.

Liang, Percy and Christopher Potts. 2015. Bringing machine learning and compositional semantics together. *Annu. Rev. Linguist.*, 1(1):355–376.

Lieto, Antonio, Antonio Chella, and Marcello Frixione. 2017. Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically inspired cognitive architectures*, 19:1–9.

Linzen, Tal and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Liu, Chaoming, Wenhao Zhu, Xiaoyu Zhang, and Qiuhong Zhai. 2023. Sentence part-enhanced bert with respect to downstream tasks. *Complex & Intelligent Systems*, 9(1):463–474.

Loula, João, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114.

Löhr, Guido. 2017. Abstract concepts, compositionality, and the contextualism-invariantism debate. *Philosophical Psychology*, 30(6):689–710.

Manning, Christopher D, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, ACL (Association for Computational Linguistics).

Martin, Andrea E and Giosuè Baggio. 2020. Modelling meaning composition from formalism to mechanism.

Martin, Andrea E and Leonidas AA Doumas. 2020. Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B*, 375(1791):20190306.

McClelland, James L, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974.

McCoy, R Thomas, Junghyun Min, and Tal Linzen. 2020. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227.

Michalon, Olivier, Corentin Ribeyre, Marie Candito, and Alexis Nasr. 2016. Deeper syntax for better semantic parsing. In *Coling 2016-26th International Conference on*

Computational Linguistics.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Montague, Richard. 1970. Universal grammar. *Theoria*, 36:373–398.

Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Niven, Timothy and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664.

Ontanon, Santiago, Joshua Ainslie, Zachary Fisher, and Vaclav Cvicek. 2022. Making transformers solve compositional tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3591–3607.

Opitz, Juri, Angel Daza, and Anette Frank. 2021a. Weisfeiler-leman in the bamboo: Novel amr graph metrics and a benchmark for amr graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.

Opitz, Juri, Angel Daza, and Anette Frank. 2021b. Weisfeiler-leman in the bamboo: Novel amr graph metrics and a benchmark for amr graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.

Opitz, Juri and Anette Frank. 2022a. Sbert studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 625–638.

Opitz, Juri and Anette Frank. 2022b. Sbert studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for*

Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 625–638.

O'shea, James, Zuhair Bandar, and Keeley Crockett. 2014. A new benchmark dataset with production methodology for short text semantic similarity algorithms. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4):1–63.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pagin, Peter and Dag Westerståhl. 2010. Compositionality i: Definitions and variants. *Philosophy Compass*, 5(3):250–264.

Palmer, Martha, Claire Bonial, and Jena D Hwang. 2016. Verbnet: Capturing english verb behavior, meaning, and usage. *The Oxford handbook of cognitive science*, page 315.

Patel, Arkil, Satwik Bhattamishra, Phil Blunsom, and Navin Goyal. 2022. Revisiting the compositional generalization abilities of neural sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 424–434.

Pavlick, Ellie. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Pelletier, Francis Jeffry. 2017. Compositionality and concepts—a perspective from formal semantics and philosophy of language. *Compositionality and Concepts in Linguistics and Psychology*, page 31.

Philipp, Markus, Tim Graf, Franziska Kretzschmar, and Beatrice Primus. 2017. Beyond verb meaning: Experimental evidence for incremental processing of semantic roles and event structure. *Frontiers in Psychology*, 8:1806.

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 4609–4622.

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Ranasinghe, Tharindu, Constantin Orăsan, and Ruslan Mitkov. 2019. Enhancing unsupervised sentence similarity methods with deep contextualised word representations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 994–1003.

Reimers, Nils. 2021. sentence-transformers.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020a. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020b. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Association for Computational Linguistics, Online.

Rossi, Andrea, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–49.

Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983.

Shaw, Peter, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Association for Computational Linguistics, Online.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227, PMLR.

Shi, Haoyue, Hao Zhou, Jiaze Chen, and Lei Li. 2018. On tree-based neural sentence modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4631–4641.

Simoulin, Antoine and Benoît Crabbé. 2022. Unifying parsing and tree-structured models for generating sentence semantic representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 267–276.

Smolensky, Paul, Richard McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. 2022. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems. *AI Magazine*, 43(3):308–322.

Socher, Richard, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211.

Soulos, Paul, Tom McCoy, Tal Linzen, and Paul Smolensky. 2019. Discovering the compositional structure of vector representations with role learning networks. *arXiv preprint arXiv:1910.09113*.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Stewart, Terrence and Chris Eliasmith. 2009. Compositionality and biologically plausible models.

Su, Jianlin, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Sun, Yu, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI conference on artificial intelligence*, 34(05):8968–8975.

Szabó, Zoltán Gendler. 2012. The case for compositionality.

Szabó, Zoltán Gendler. 2020. Compositionality. *The Stanford Encyclopedia of Philosophy*.

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546.

Tripathy, Jatin Karthik, Sibi Chakkaravarthy Sethuraman, Meenalosini Vimal Cruz, Anupama Namburu, P Mangalraj, Vaidehi Vijayakumar, et al. 2021. Comprehensive analysis of embeddings and pre-training in nlp. *Computer Science Review*, 42:100433.

Tsukagoshi, Hayato, Ryohei Sasano, and Koichi Takeda. 2021. Defsent: Sentence embeddings using definition sentences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–418.

Vankov, Ivan I. and Jeffrey S. Bowers. 2019. Training neural networks to encode symbols enables combinatorial generalization. *Philosophical Transactions of the Royal Society B*, 375.

Vankov, Ivan I and Jeffrey S Bowers. 2020. Training neural networks to encode symbols enables combinatorial generalization. *Philosophical Transactions of the Royal Society B*, 375(1791):20190309.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vecchi, Eva M, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, 41(1):102–136.

Wang, Bin, C-C Jay Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8:e19.

Wang, Jiongxiao, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*.

Wang, Tiexin, Hui Shi, Wenjing Liu, and Xinhua Yan. 2022. A joint framenet and element focusing sentence-bert method of sentence similarity computation. *Expert Systems with Applications*, 200:117084.

Wang, Zhiguo, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1340–1349.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Webson, Albert and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344.

Wei, Chengwei, Bin Wang, and C-C Jay Kuo. 2023. Synwmd: Syntax-aware word mover's distance for sentence similarity evaluation. *Pattern Recognition Letters*, 170:48–55.

Westerståhl, Dag. 1998. On mathematical proofs of the vacuity of compositionality. *Linguistics and Philosophy*, pages 635–643.

Yao, Yuekun and Alexander Koller. 2022. Structural generalization is hard for sequence-to-sequence models. *arXiv preprint arXiv:2210.13050*.

51

Young, Tom, Devamanyu Hazarika,
Soujanya Poria, and Erik Cambria. 2018.
Recent trends in deep learning based
natural language processing. *ieee
Computational intelligenCe magazine*,
13(3):55–75.

Yu, Changlong, Tianyi Xiao, Lingpeng Kong,
Yangqiu Song, and Wilfred Ng. 2022. An
empirical revisiting of linguistic
knowledge fusion in language
understanding tasks. *arXiv preprint
arXiv:2210.13002*.

Yu, Lang and Allyson Ettinger. 2020.
Assessing phrasal representation and
composition in transformers. In
*Proceedings of the 2020 Conference on
Empirical Methods in Natural Language
Processing (EMNLP)*, pages 4896–4907.

Žabokrtský, Zdeněk, Daniel Zeman, and
Magda Ševčíková. 2020. Sentence meaning
representations across languages: what
can we learn from existing frameworks?
Computational Linguistics, 46(3):605–665.

Zadrozny, Wlodek. 1994. From
compositional to systematic semantics.
Linguistics and philosophy, 17:329–342.

Zhang, Honghua, Liunian Harold Li, Tao
Meng, Kai-Wei Chang, and Guy Van den
Broeck. 2022. On the paradox of learning
to reason from data. *arXiv preprint
arXiv:2205.11502*.

Zhang, MeiShan. 2020. A survey of
syntactic-semantic parsing based on
constituent and dependency structures.
Science China Technological Sciences,
63(10):1898–1920.

Zhang, Zhuosheng, Yuwei Wu, Hai Zhao,
Zuchao Li, Shuailiang Zhang, Xi Zhou,
and Xiang Zhou. 2020a. Semantics-aware
bert for language understanding. In
*Proceedings of the AAAI Conference on
Artificial Intelligence*, volume 34, pages
9628–9635.

Zhang, Zhuosheng, Yuwei Wu, Hai Zhao,
Zuchao Li, Shuailiang Zhang, Xi Zhou,
and Xiang Zhou. 2020b. Semantics-aware
bert for language understanding.
*Proceedings of the AAAI Conference on
Artificial Intelligence*, 34(05):9628–9635.

Zhou, Denny, Nathanael Schärli, Le Hou,
Jason Wei, Nathan Scales, Xuezhi Wang,
Dale Schuurmans, Olivier Bousquet, Quoc
Le, and Ed Chi. 2022. Least-to-most
prompting enables complex reasoning in
large language models. *arXiv preprint
arXiv:2205.10625*.