How Is a "Kitchen Chair" like a "Farm Horse"? Exploring the Representation of Noun-Noun Compound Semantics in Transformer-based Language Models

Mark Ormerod
Queen's University Belfast
mormerod01@qub.ac.uk

Jesús Martínez del Rincón
Queen's University Belfast

Barry Devereux
Queen's University Belfast

Despite the success of Transformer-based language models in a wide variety of natural language processing tasks, our understanding of how these models process a given input in order to represent task-relevant information remains incomplete. In this work, we focus on semantic composition and examine how Transformer-based language models represent semantic information related to the meaning of English noun-noun compounds. We probe Transformer-based language models for their knowledge of the thematic relations that link the head nouns and modifier words of compounds (e.g., KITCHEN CHAIR*: a chair* located in *a kitchen). Firstly, using a dataset featuring groups of compounds with shared lexical or semantic features, we find that token representations of six Transformer-based language models distinguish between pairs of compounds based on whether they use the same thematic relation. Secondly, we utilize fine-grained vector representations of compound semantics derived from human annotations, and find that token vectors from several models elicit a strong signal of the semantic relations used in the compounds. In a novel "compositional probe" setting, where we compare the semantic relation signal in mean-pooled token vectors of compounds to mean-pooled token vectors when the two constituent words appear in separate sentences, we find that the Transformer-based language models that best represent the semantics of noun-noun compounds also do so substantially better than in the control condition where the two constituent works are processed separately. Overall, our results shed light on the ability of Transformer-based language models to support compositional semantic processes in representing the meaning of noun-noun compounds.*

Action Editor: Kevin Duh. Submission received: 21 September 2022; revised version received: 29 April 2023; accepted for publication: 17 June 2023.

https://doi.org/10.1162/coli_a_00495

1. Introduction

One rapidly growing strand of Natural Language Processing (NLP) research is that of determining whether neural language models encode certain linguistic properties, and, if so, understanding how these properties are represented. This goal has brought together a variety of researchers and analysis techniques from fields such as machine learning, linguistics, psychology, and neuroscience (Linzen 2019; Abnar et al. 2019; Gauthier and Levy 2019; Anderson et al. 2021). With the advent of Transformer-based language models (Vaswani et al. 2017) such as BERT (Devlin et al. 2018), there has been a surge of interpretability research on this type of architecture (in particular, the study of BERT and related models has gained much popularity in a wave of research sometimes referred to as "BERTology" [Rogers, Kovaleva, and Rumshisky 2021]). To date, however, most of the research into the interpretation of Transformer-based language models has focused on their syntactic knowledge, and while there have been investigations into their semantic capabilities (e.g., Ettinger 2020; Tenney, Das, and Pavlick 2019), our understanding of how Transformers process semantic information remains largely incomplete. In contrast to syntax, where explicit representations of the grammatical structures of interest are available, a challenge faced in probing Transformer-based language models for semantics is finding suitable experimental frameworks for investigating processes relating to semantic representation and semantic composition.

In this work, we examine the extent to which Transformer-based language models have implicit knowledge of the thematic relations used in noun-noun compounds and explore how this information is encoded in the intermediary vector representations of these models. To this end, we perform layer-wise representational analysis on six different types of Transformer-based language models, covering a range of training objectives, training data, and total number of parameters.

1.1 Noun-noun Compounds and Semantic Composition

Noun-noun compounds are simple two-word phrases made up of a head noun that is modified by a modifier word. For example:

1. PUBLIC HOUSE
2. BRICK HOUSE
3. COUNTRY HOUSE

Despite their simple and consistent syntax, the meaning of the two words in a compound can combine to form a meaning for the phrase as a whole in semantically diverse ways. An interesting feature of noun-noun compounds is that, despite the semantic relation between the head noun and modifier word not being explicitly present in the phrase, their meaning is usually completely transparent to humans, even when the compound is a novel construction (van Jaarsveld and Rattink 1988). Following linguistic analysis of such compounds, we can describe their meaning with a taxonomy of *thematic relations*; that is, PUBLIC HOUSE is a house *for* the public, a BRICK HOUSE is a house *made of* brick, and a COUNTRY HOUSE is a house *located in* the country (Levi 1978; Gagné and Shoben 1997). Other approaches to compound taxonomies include Lees Robert (1960)

(an early proponent of the idea that there are a fixed number of relations for a particular head-modifier word combination), Downing (1977) (who in contrast emphasizes that the relation that describes a compound can take on any interpretation and is determined pragmatically), and more recent work such as Tratz and Hovy (2010), who create a novel taxonomic inventory by integrating several previous schemes. An alternative approach to using a taxonomy of thematic relations to interpret compounds is the dimension-based approach (Murphy 1988), which views the head noun as a schema defining a set of dimensions, each with a set of possible values. In this view, the modifier word will then fill one of these dimensions during the process of conceptual combination (Gagné and Shoben 1997). The ability of a computational model to relate the features of the two constituent words of the larger expression such that the properties of the resulting compound representation correlate with human judgment values would constitute a demonstration of semantic compositionality (Mitchell and Lapata 2010), although we would not expect the operations that would enable such a process in neural networks to be encoded in a set of systematic rules (Baroni 2020). In any case, we consider noun-noun compounds to be well-suited for investigating semantic representation and conceptual composition in Transformer-based language models—indeed, in the psycholinguistics literature, the interpretation of noun-noun compounds has proven to be a lively research area for both theories of concept representation and conceptual composition (Gagné and Shoben 1997; Murphy 2002; Estes and Hasson 2004; Devereux and Costello 2012; Lynott and Connell 2010; Maguire et al. 2007; Westerlund and Pylkkänen 2017).

In the NLP context, work on the computational interpretation of noun-noun compounds has involved classifying noun-noun compound semantic relations using a variety of features, such as semantic class information and various syntactic features (Girju et al. 2004), and lexical similarity and co-occurrence information (Ó Séaghdha and Copestake 2007; Devereux and Costello 2005). Subsequent work by Tratz and Hovy (2010) used surface features of word forms for automatically interpreting noun-noun compounds. Work by Reddy, McCarthy, and Manandhar (2011) found evidence that distributional word-space models can predict human compositionality judgments of noun-noun compounds. Other innovations in noun-noun compound interpretation include utilizing paraphrase models (Shwartz and Dagan 2018), using transfer learning and multi-task learning (Fares, Oepen, and Velldal 2018), or framing this problem as a verb paraphrasing task (Nakov 2019). More recently, Shwartz and Dagan (2019) demonstrated the power of using contextualized word embeddings (including Transformer representations) for noun-noun compound relation classification. While previous authors have generally aimed at using state-of-the-art NLP models and machine learning techniques to explore the limits of noun-noun compound relation classification, we use noun-noun compounds as a means of interpreting how Transformer-based language models build representations of the semantic relationships that exist between the constituent words.

1.2 Representational Similarity Analysis

Our analyses make use of Representational Similarity Analysis (RSA), a multivariate statistical methodology first developed in imaging neuroscience (Kriegeskorte, Mur, and Bandettini 2008). RSA allows for the comparison of different kinds of multivariate data, enabling us to compare representational vectors with both different dimensionalities (e.g., comparing two models with different hidden vector sizes) and wholly

disparate modes of representation (e.g., comparing language model token vectors to a set of linguistic features). In order to compare two sets of n representations, we first construct a representational dissimilarity matrix (RDM) for each of the two models that captures the pair-wise dissimilarity between all of the stimuli (typically computed as $1-$ the Pearson's correlation between each of the representations), producing one $n \times n$ matrix for each of the two models. We can then take a second-order correlation between the two RDMs to measure the similarity between the two models' internal dissimilarity structure given our set of stimuli. This approach to representational analysis has the advantage of capturing patterns in distributed information that may not necessarily be encoded in a particular dimension of a token vector (Nili et al. 2014). A broad overview of how RSA is integrated into our analysis pipeline is given in Figure 1.

1.3 Research Questions and Predictions

In this work we target two primary theoretical research questions:

1. **To what extent do Transformer-based language models encode noun-noun compound relational semantics?** We consider whether token vectors in Transformer models can broadly distinguish between semantic classes of noun-noun compounds (e.g., *H made of M* versus *H located in M*), and whether we can recover fine-grained information about all possible relations between the head noun (*H*) and modifier (*M*) (as informed by human judgments of the possible semantic interpretations of the compound).
2. **How is thematic relation information encoded in Transformer model representations?** If Transformer models can to some extent represent relational semantics between the head and modifier noun, we wish to understand how this information is encoded within the token vectors of the model. In particular, we identify the three following areas of investigation: (1) whether this relational representation relies on memorizing distributional co-occurrence information (as opposed to a step-wise dynamic process where head and modifier nouns are contextually composed and relational information gradually emerges), (2) whether this information is localized within a particular token span within the compound (i.e., in the head or modifier token vectors, as opposed to a broader context), and (3) whether this information is localized to a particular layer or set of layers.

Given the growing body of research that demonstrates the ability of such models to encode rich linguistic information on natural language input, we expect that English and multilingual Transformer models would be able to produce relation representations that broadly distinguish between classes of English noun-noun compounds. Additionally, we also predict that these models can to some extent represent fine-grained relational-semantic information about noun-noun compounds such that they align with human ratings of the detailed and multifaceted relationship between the head and modifier noun, but that this fine-grained knowledge may vary across model architectures.

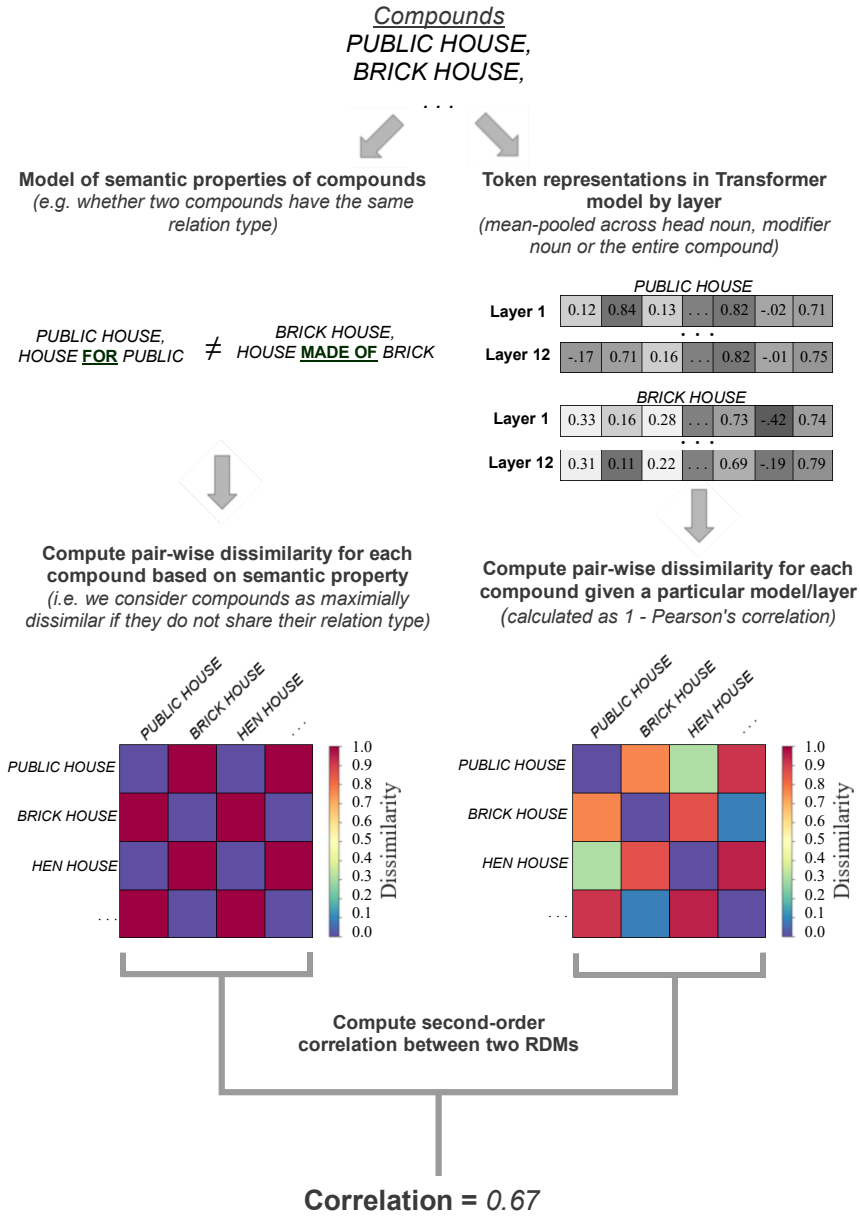


Figure 1
Overview of our feature extraction and Representational Similarity Analysis pipeline. Here we show the procedure for calculating the RDM for the Relation Category experiments (Experiments 1a and 1b), where compounds are counted as maximally dissimilar if they do not share a primary thematic relation. The second-order correlation between the RDMs measures the degree to which the given model of compound semantics is reflected in the Transformer representations.

The question of the extent to which Transformer models can perform compositional tasks is a contested area within the interpretability literature (Ontanon et al. 2022). While there are several studies that demonstrate the compositional ability of Transformer models in controlled settings (Murty et al. 2022; Csordás, Irie, and Schmidhuber 2021), other work has called into question the ability of these models to achieve nuanced semantic composition (Yu and Ettinger 2021), or have suggested that the success of such models to interpret compounds may depend on the memorization of token distribution information (Li, Carlson, and Potts 2022; Coil and Shwartz 2023). In the most relevant work to the present study, Yu and Ettinger (2020) found little evidence of compositional semantics (using a dataset of two-word phrases) in a range of Transformer models. Given this body of literature, we predict that the Transformer model that best encodes relational information will be able to compose head and modifier words to produce representations that capture broad relational information (as opposed to rich fine-grained information about additional facets of the head-modifier semantic relation). With respect to the question of whether relational information will be localized in the head noun tokens or modifier noun tokens (or distributed across several words), we are interested in relating the interpretation of noun-noun compounds in Transformer models to the psychological literature, which suggests that the ease of interpretation of a compound is predicted by the association of the modifier word (but not the head word) with the relation type (Gagné 2001; Devereux and Costello 2006). Nevertheless, we expect that this information will always be to some extent distributed over both the head and the modifier word, as the attention mechanism in Transformer models allows information to flow from each token vector in a particular layer to all token vectors in the subsequent layer. With respect to the question of where the relational information will be localized, we expect that such semantic information will be encoded in later layers, following work such as Tenney, Das, and Pavlick (2019) that shows that high-level semantic information typically surfaces in later layers of Transformer models.

1.4 Contributions

We find that all layers of the four monolingual English-language models produce representations of compound relations that more strongly correlate with human semantic judgments when head and modifier nouns are concurrently processed as a compound, compared with the baseline multilingual and Japanese models. To our knowledge, these experiments are the first to show that Transformer-based language models meaningfully encode implicit relational semantic knowledge about the meaning of noun-noun compounds. Across the series of experiments, the results suggest that the Transformer-based language models that encode the strongest representation of thematic relations dynamically integrate their knowledge of the intrinsic properties of the head and modifier concepts in order to encode the semantic relationship between these words, rather than only relying on the lexical information of the component words in isolation, or information about concept-relation frequency.

2. Materials

We use two datasets to explore the representation of English noun-noun compounds in Transformer-based language models, covering 30 Transformer models across 6 types of models (including 25 instantiations of the same Transformer model trained with different randomized weight initializations).

2.1 Data

The first dataset (Gagné 2001) is made of 300 English noun-noun compounds that are organized into 60 groups of five compounds each. The dataset was originally compiled in order to investigate lexical and relational priming in psycholinguistic experiments on human understanding of noun-noun compounds. Using a taxonomy of 16 thematic relation types, each compound is annotated with the most appropriate thematic relation for describing the semantic relationship that exists between the head noun and the modifier (Gagné and Shoben 1997). Each group of five compounds is composed of a target compound followed by four compounds that feature (a) either a different head or modifier word from the target compound and (b) either the same or a different relation between the head noun and modifier word from the target compound, covering four different experimental conditions (see Table 1). Within each group, each modifier and head occurs with a thematic relation that is highly frequent with the modifier (e.g., the modifier MOUNTAIN often occurs with a *located in* relation, as in MOUNTAIN BREEZE but rarely occurs with an *about* relation, as in MOUNTAIN MAGAZINE). In this way, the occurrence of relation type with the individual modifier and head nouns is controlled in the experimental design (see Gagné [2001] for details).

The taxonomy of 16 thematic relation types utilized by Gagné (2001) is a useful, but rather coarse-grained, representation of the semantics of the relation instantiated for particular compounds. In many cases, several relation types may capture the meaning of a given compound, to varying degrees. We therefore also make use of a dataset of 60 compounds (a subset of the 300 compounds described above) where 34 participants rated the appropriateness of 18 different thematic relations for every compound (Devereux and Costello 2005).

These relation types and the number of total mentions for each of the types are presented in Figure 2. Compounds for which the semantic link between the head word and modifier word are similar (e.g., PROPANE STOVES and GAS LAMPS) tend to have similar distributions of appropriateness ratings across the thematic relations (see Devereux and Costello [2005] for details), and the thematic relation ratings can therefore be utilized as 18-dimensional vector representations of the relational meaning used in compounds. Relation vectors for three compounds are presented in Figure 3. Here we observe that PROPANE STOVES and GAS LAMPS are close together in the relation space (consistent with the "H uses M as fuel" relationship found in both compounds), whereas PROPANE STOVES and RAIN DROPS have very different relation vectors.

For all compounds in the datasets described above, we construct a corpus of simple, neutral sentences in the form of "It is a {*compound*}" for singular compounds (e.g.,

Table 1
Five compounds that make up one of the 60 compound groups in the Gagné (2001) noun-noun compound dataset, used in our Relation Category RSA experiments.

Modifier (M)	Head (H)	Experimental condition	Thematic relation
mountain	breeze	Target	*H LOCATED M*
kitchen	breeze	Same head noun, same relation	*H LOCATED M*
storm	breeze	Same head noun, different relation	*H DURING M*
mountain	cabin	Same modifier, same relation	*H LOCATED M*
mountain	magazine	Same modifier, different relation	*H ABOUT M*

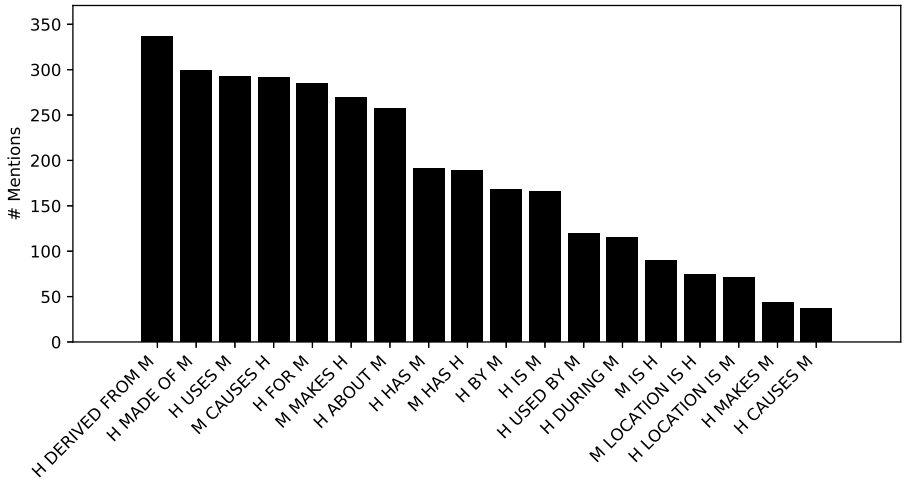


Figure 2
Distribution of relation types across the 60 compounds (Devereux and Costello 2005) dataset.

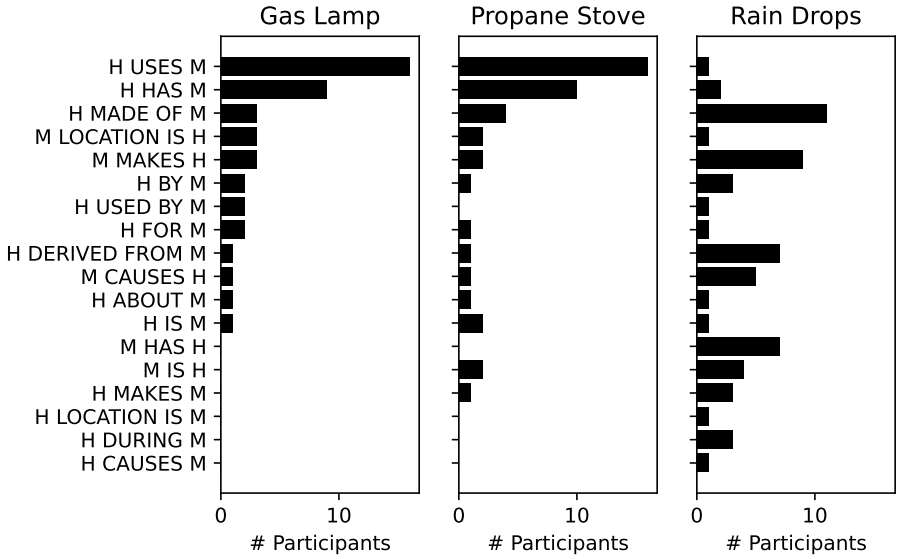


Figure 3
Sample relation vectors, based on the number of participants selecting each relation for each compound, for three compounds given in the Devereux and Costello (2005) dataset. GAS LAMP and PROPANE STOVE are close in this 18-dimensional space, reflecting the semantic similarity of the relational link (both stoves and lamps contain gas/propane that they use as fuel), while both compounds share relatively little overlap with RAIN DROPS.

"It is a wood stove"), "It is {*compound*}" for mass noun compounds (e.g., "It is solar power") and "They are{*compound*}" for plural compounds (e.g., "They are summer clothes"). The motivation for using compounds in such minimalistic sentences was to present the language models with naturalistic sentences as they would encounter

in training, while at the same time minimizing variability due to extrinsic context. As all compounds from the Devereux and Costello (2005) relation vector dataset are found in the 300 compound dataset, we use these sentences in both data settings. For our compositional analyses (the Relation Category and Processing RSA experiment [Section 3.3], the Relation Vector and Processing RSA experiment [Section 3.5], and the Compositional Probe experiment [Section 3.6]), we construct two similar sentences for all compounds using the head or modifier nouns in isolation; for example, the compound WAR RIOTS yields the sentences "It is a war" and "They are riots". After creating the corpus of minimalistic sentences, we wished to evaluate whether these constructions are particularly implausible (which could limit the generalizability of our findings) and check whether there are large disparities in sentence plausibility across different compound relations (which could potentially introduce a confound in our analysis). To this end, we used GPT-2 (Radford et al. 2019) (an autoregressive Transformer model) to calculate the perplexity of each sentence before measuring whether perplexity significantly correlated with relation magnitude for any of the relation vector dimensions. We measured an average perplexity of 267.12, compared to an average perplexity of 774.51 across sentences of similar lengths (i.e., between 5 and 14 words long) in the WikiText-2 dataset (Merity et al. 2016).[1] We then measured the Pearson's correlation between the perplexity of each sentence and the magnitude of the relation for each of the 18 dimensions in the relation vector, finding no significant correlation between any relation type and the likelihood of sentences in our corpus.

2.2 Models

In this work we considered six different Transformer-based language models (Vaswani et al. 2017). We follow the BlackboxNLP 2020 Shared Interpretation Mission (Alishahi et al. 2020) in the choice of models: BERT (bert-base-cased)[2] (Devlin et al. 2018), BERT-Japanese (bert-base-japanese)[3], RoBERTa (roberta-base) (Liu et al. 2019), DistilRoBERTa (disilroberta-base) (Sanh et al. 2020), XLM (xlm-mlm-xnli15-1024) (Lample and Conneau 2019), and XLNet (xlnet-base-cased) (Yang et al. 2019). All of the Transformer models we target are masked language models (although *xlnet-base-cased* has been exposed to an autoregressive pre-training regime). For our BERT analysis we have made use of the MultiBERTs resource (Sellam et al. 2021), which enables us to carry out experiments on 25 different versions of the *bert-base-uncased* model that have been trained with different starting weight initializations and different shuffling of the training data, allowing for a more robust analysis of the representational trends in BERT-style models. These models were chosen to assess whether our analyses generalize over a diverse range of Transformer-based language model design choices, including monolingual/multilingual data, model size, and choice of training objective and training data. Furthermore, these models have also been used by the most relevant studies to the current work (Yu and Ettinger 2020, 2021) enabling a more direct comparison between our approach to probing Transformer models for compositional semantics and theirs.

1 Perplexity was calculated for each sentence separately before taking the average. When perplexity is calculated using the common sliding-window strategy (using preceding sentences to predict tokens), the average perplexity is 25.17 across the WikiText-2 dataset. In the sliding-window setting the model is able to leverage a large amount of contextual information to predict the recurring template structure, which would produce an extremely low perplexity score for our generated sentences.

2 Model names in the HuggingFace library are given in parentheses.

3 `https://github.com/cl-tohoku/bert-japanese`.

A priori, we hypothesize that the English-specific models should be best at representing the relation semantics of English noun-noun compounds, and the BERT-Japanese model is primarily included as a control. In all experiments, we do not fine-tune the models, as we aim to evaluate each model's capacity for semantic relation representation given its standard training on a general domain language modeling task. All models were accessed using the Hugging Face library (Wolf et al. 2020).

3. Experiments and Results

We design a variety of experiments in order to assess whether the six Transformer-based language models encode the semantic relations between the head and modifier words in English noun-noun compounds in their token vector representations. The design of these experiments focuses on examining differences across models, across layers, and across the constituent words of the compounds. In the Relation Category and Relation Vector experiments (Sections 3.2, 3.3, 3.4, and 3.5), we use RSA (Kriegeskorte, Mur, and Bandettini 2008) to measure the degree to which patterns of activation in the models reflect the thematic relation information corresponding to the interpretation of compounds. In the RSA analyses, we consider both a "course-grained" representation of thematic relation information, where similarity is based on whether the thematic relation taxonomic label is the same or different across compounds (Figure 1), and a "fine-grained" representation, where a measure of pairwise similarity between relation vectors in relation space is used to capture similarity of relational meaning. In the Compositional Probe experiment (Section 3.6), we use linear regression probing to measure the decodability of the fine-grained relation vectors given different data ablation conditions. In our experiments we provide the model with a minimal sentence containing a noun-noun compound, and extract mean-pooled token vector representations across particular token spans at each layer. For the Relation Category RSA (Section 3.2) and the Relation Vector RSA (Section 3.4) experiments, we mean-pool across (1) tokens in the head noun, (2) tokens in the modifier noun, and (3) all of the tokens in both the head and modifier noun. In other experiments we only consider mean-pooled representations of tokens in the entire compound. In all of our results figures, we report an average value for the MultiBERTs models and show the standard error over the range of results as an error bar. Significant differences for the MultiBERTs models in the Relation Category and Processing Condition experiment, the Relation Vector and Processing Condition experiment, and the Compositional Probe experiment were checked using paired t-tests.

3.1 Experimental Design and Controls

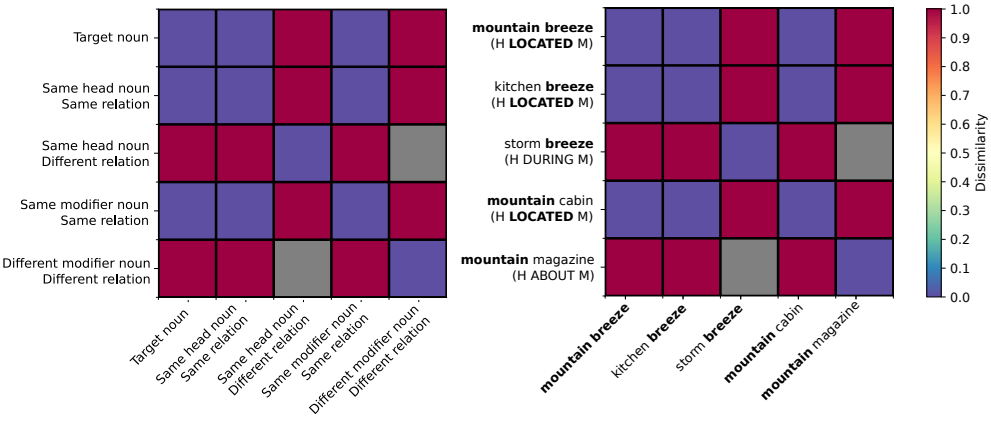
In any analysis of whether and how a particular aspect of linguistic knowledge is encoded in a language model, a key consideration is whether the analysis is sensitive to experimental confounds and other spurious cues that correlate with the phenomena of interest (Yu and Ettinger 2020). In the case of analyzing models for semantic composition, a particular issue is the potential correlation between the lexical forms and the relational information describing the semantics of composition (for example, the compounds MOUNTAIN STREAM and MOUNTAIN CABIN both use a *located in* thematic relation, but they also both contain the modifier MOUNTAIN). In this work, therefore, we make use of three types of experimental control, in order to separate semantic composition from the representation of lexical information. Firstly, we make use of a psycholinguistic experimental design, in which the thematic relations used in the analyzed

compounds are counterbalanced with the modifier and head words appearing in the compounds. Secondly, we include a multilingual language model and a Japanese language model as controls in the analysis, on the hypothesis that such models, compared to English-language models, will not adequately represent the compositional meaning of English noun-noun compounds even if they are sensitive to word overlap across compounds. Finally, we also construct a novel "compositional probe" that measures the difference in semantic relation representation when a compound is processed in a single sentence versus when the head and modifier nouns are processed in separate sentences.

3.2 Experiment 1a: Relation Category RSA

3.2.1 Overview. In the Relation Category RSA experiment we use RSA to investigate whether representations extracted from the Transformer-based language models distinguish between noun-noun compounds based on whether pairs of compounds share the same thematic relation type. For this experiment we use the Gagné (2001) 300 compound relation group dataset. We only consider compound pairs within each of the 60 groups, following the experimental design of the Gagné (2001) study. The 5×5 RDM for each group encodes whether the same or different thematic relation is used for each pair of compounds in the group (Figure 4). As two of the compounds in each group are marked only as differing in thematic relation from the target compound (e.g., the STORM BREEZE – MOUNTAIN MAGAZINE pair in Figure 4), we do not include this pair of compounds, as these experimental conditions are not compared in the Gagné (2001) experimental design.

In the Relation Category RSA experiment, we present sentences to the model that contain each compound (e.g., "They are war riots"). The data for the experimental



a) RDM structure for one compound group, labelled according to categorical distinctions with respect to the target noun.

b) RDM for a particular compound group with the target noun **mountain breeze.** Relations, head and modifier words in bold are shared with the target noun.

Figure 4
Representational Dissimilarity Matrix (RDM) for same/different thematic relation taxonomic category for one compound group in the Gagné (2001) dataset. The 300 experimental items are divided into 60 groups of noun-noun compounds with this similarity structure. In our experiments, we ignore the relation pairs marked in gray as we have no ground-truth similarity information for this compound pair (these two compounds are classified as not having the same relation as the primary compound of the group (i.e., *mountain breeze*) and as such may or may not differ between each other).

RDMs that we consider is the mean-pooled token spans for the tokens that comprise (1) the modifier word, (2) the head noun, and (3) the whole compound. We construct three experimental RDMs for each layer of each model by taking the cosine similarity between all pairs of noun-noun compounds for the three choices of representation. For each 5×5 compound group RDM, we measure the second-order similarity between each experimental RDM and the ground-truth RDM using Pearson's correlation (with the correlation restricted to the upper-triangular part of the matrix, as is standard in RSA). The strength of this second-order correlation reflects the degree to which the pattern-information of the model activation vectors reflects the representational content encoded by the ground-truth RDM (in this case, the identity of the thematic relation category used in each compound). We report the average correlation across all 60 compound groups for each layer of each model. This design allows us to measure the relative strength of the coarse-grained thematic relation signal across a variety of different models, layers, representation types, thematic relations, and compounds.

3.2.2 Results. The results for the Relation Vector RSA experiment are given in Figure 5. Overall, we generally see positive correlations between the ground-truth RDMs and

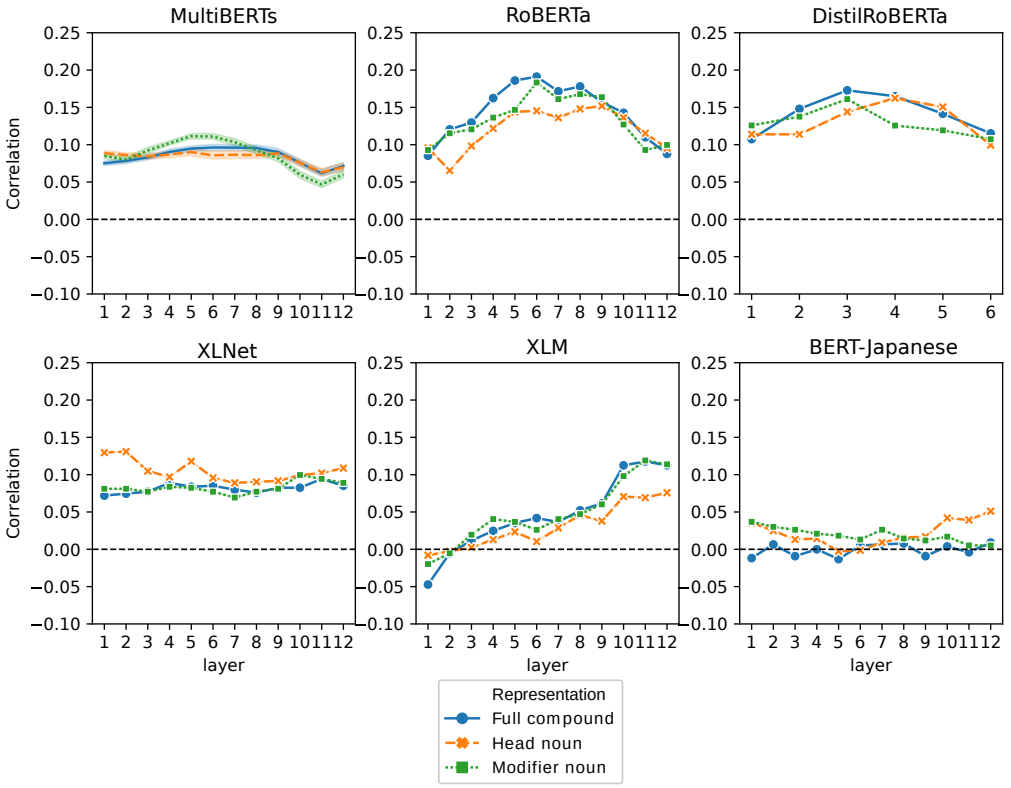


Figure 5
Results of the Relation Category RSA experiment (Section 3.2). Average correlation between the same thematic relation ground-truth RDM and experimental RDMs constructed using mean-pooled token-span representations for 6 types of Transformer-based language models (300 sentences, correlation averaged across 60 compound groups).

the activation RDMs; in particular, the four English-language models consistently show relatively high correlations across all layers, indicating that they are sensitive to the representational geometry captured by our ground-truth thematic relation category RDM. The two models that are not English monolingual models (*bert-base-japanese* and *xlm-mlm-xnli15-1024*) achieve lower correlations than the other four models, although *xlm-mlm-xnli15-1024* begins to produce more strongly correlated representations of the coarse-grained thematic relation signal in its final few layers. While *bert-base-japanese* acts as a control model (reflected in its relatively low correlation overall), this model was still able to consistently achieve correlations greater than zero using head-noun and modifier noun representations. This result may suggest that the relatively small amount of English language content encountered in the *bert-base-japanese* training data (i.e., Japanese Wikipedia articles) could be sufficient to learn to represent at least some information about English semantic categories. However, the mean-pooled compound representation does not tend to correlate much more strongly with the coarse-grained thematic relation RDM than the modifier and head representations for the majority of the layers. This may suggest that at a given layer the head and modifier token representations encode more information relevant to the primary thematic relation than does the entire compound, or that the mean-pooling approach does not preserve the representational pattern that distinguishes compounds by the semantics of their primary relation. One notable trend is that the overall best representation for all of the monolingual English models is either the head-noun or the mean-pooled compound. These correlations occur in the early-middle layers of the model, while *bert-base-japanese* and *xlm-mlm-xnli15-1024* produce their best representations of the coarse-grained thematic relation RDM in their final layers of processing. In particular, *xlm-mlm-xnli15-1024* shows an almost monotonic increase in correlation into later layers, clearly indicating that this model best represents thematic relation information for compounds in the final three layers of processing.

Despite observing a clear disparity between the correlation strengths of the baseline Japanese model and the other five types of Transformer models, we note that the overall effect sizes are not particularly high, peaking at a moderately positive correlation of around 0.2 for the *roberta-base* representations. One reason for this range of effect strengths is that the token vectors of these Transformer models encapsulate much more information about the compound nouns (broader semantic and syntactic information, world knowledge, etc.) than the relational information alone and as such there will be a limit on the amount of variance between the representations that is explained by the relation category only. This underlines the importance of including a baseline model in order to contextualize the strength of alignment between the relation category dissimilarity and dissimilarity patterns measured within a given model.

3.2.3 Summary. We found that excluding early layers of *xlm-mlm-xnli15-1024* and most layers of the baseline *bert-base-japanese* model, all models produced representations that moderately positively correlated with the relation category distinction, a finding that agrees with our prediction for our first research question (that the English models would represent relational information between head and modifier words in noun-noun compounds). There are mixed results for where this information is localized (both with respect to token span and layer), but overall we find that the middle layers are more highly correlated with the relational signal for BERT-style models and there is a clear trend for later layers of *xlm-mlm-xnli15-1024* to elicit stronger correlations with the relation category RDM.

3.3 Experiment 1b: Relation Category and Processing Condition RSA

3.3.1 Overview. As mentioned in the Introduction, a variety of possible confounds exist in the analysis of noun-noun compound meaning, including the potential co-occurrence of thematic relation information with the individual constituent words. This experiment is therefore designed to measure the difference in the strength of the coarse-grained thematic relation signal when Transformer-based language models are presented with the modifier word and the head noun together in a compound phrase (e.g., *"They are war riots"*) compared to when we compose representations of a compound from the activations to the head noun and modifier word where they are processed in two separate sentences (e.g., *"It is a war"* and *"They are riots"*). If there is a greater correlation with the thematic relation RDM when the two constituent words of a compound are processed as a noun-noun compound phrase in the same sentence compared to when they are processed separately, this would indicate that the model represents the semantic information of the thematic relation in the noun-noun compound rather than only relying on information about the co-occurrence of a particular head or modifier word with a particular thematic relation category. In this experiment we use a similar RSA procedure to that of the first Relation Category RSA experiment (described in Section 3.2) by measuring the correlation between the thematic relation RDM and mean-pooled token representations for the head and modifier extracted under two processing conditions: (i) when the head and modifier nouns of a compound are processed in the same sentence, and (ii) when the head noun and modifier noun are processed in two separate sentences.

3.3.2 Results. The results for the Relation Category and Processing Condition RSA experiment are given in Figure 6. We used one-sided paired sample t-tests for each of the layers of each model to compare the correlation strengths within the 60 compound groups across the two processing conditions. Significant effects at $p \leq 0.05$ after applying a false discovery rate controlling procedure (Benjamini–Hochberg with $\alpha = 0.05$) are indicated with asterisks.

The statistical analysis shows that most layers of the *roberta-base*, *distilroberta-base* and the MultiBERTs models represent the thematic relation better in the context where the modifier and head are presented together as a compound, compared with where they are presented separately, which is as expected if the models represent the relational semantics of the noun-noun compound phrase rather than relational information associated with the two words separately. The most striking results are for the *roberta-base* and *distilroberta-base* models—when the modifier and head noun are processed together as a compound, these models have the highest overall correlations with the relation category RDM, and furthermore these correlations are significantly higher than in the separate processing case for nine of the 12 layers of *roberta-base* and all but one layer of *distilroberta-base*. In the case of the baseline *bert-base-japanese* model, there is clearly no difference in how well the thematic relation is represented across the Together and Separate conditions. In the final layers of the multilingual *xlm-mlm-xnli15-1024* model, we see a difference in average correlation between the two conditions, but this difference is not statistically significant.

We note that for the models that show the largest differences in the compound processing case compared to the separate sentence case (*roberta-base* and *distilroberta-base*), both of these models show low correlations when the modifier and head words are not processed in the same context, an effect that is strongest in their first few layers. This result suggests that models that compose representations of semantic relations between

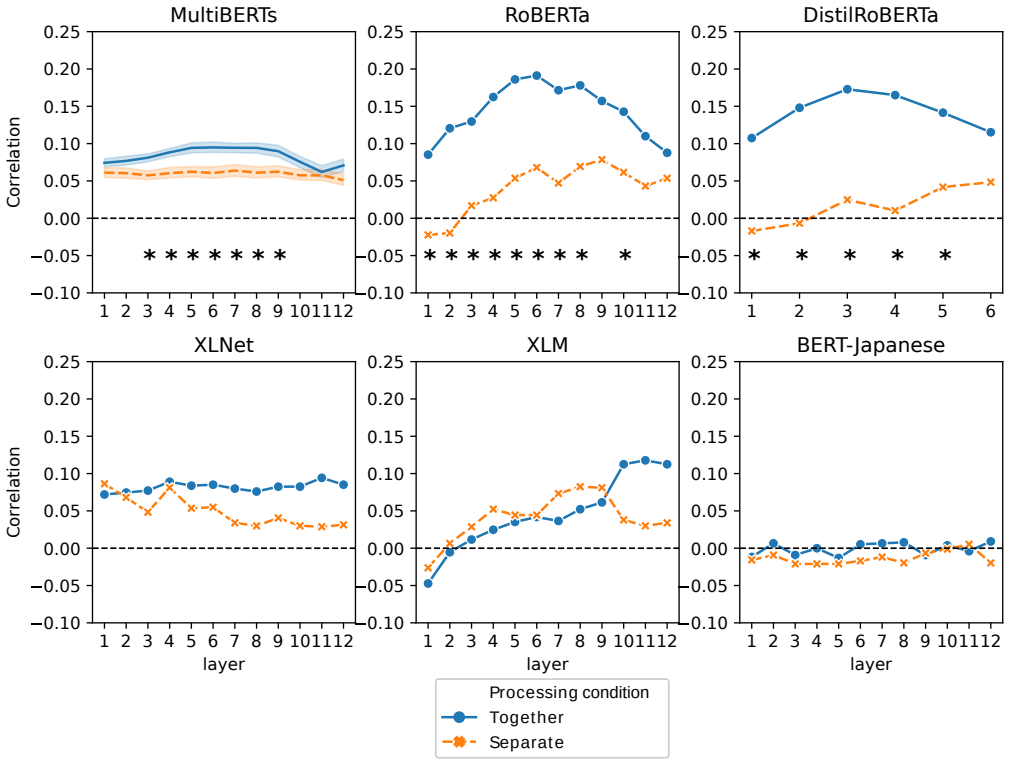


Figure 6
Results of the Relation Category and Processing Condition RSA experiment (Section 3.3). Average correlation between the same thematic relation ground-truth RDM and experimental RDMs constructed using mean-pooled token-span representations for two compound processing conditions: when the head noun and modifier noun are presented together as a noun-noun compound in the same sentence ("Together"), and when the head noun and modifier noun are presented in separate sentences ("Separate"). Results for 300 sentences; correlation averaged across 60 compound groups.

words within the same context encode information about the thematic relation more strongly than models that begin with a relatively strong association between individual word embeddings and their possible thematic relations (such as *xlnet-base-cased* and the MultiBERTs.

The first layer of *xlnet-base-cased* gives the strongest correlation to the thematic relation RDM when the words are processed separately. We can compare this result to the similar early layer bias for *xlnet-base-cased* in the Relation Category RSA experiment (Section 3.2), where the representations for the head noun in the first few layers of the model gave the strongest correlations. Taking these results together, it would appear that for representing thematic relations in noun compounds, *xlnet-base-cased* relies on distributional information of the co-occurrence between particular individual words and particular thematic relations. In particular, this information is primarily encoded in the head noun and is at its strongest closer to the embedding representation. Together with the lack of a significant difference between the Together and Separate conditions for *xlnet-base-cased*, this suggests that this model mostly relies on lexical information

about the association of words with thematic relations, rather than a truly compositional representation of the thematic relation used in compound meaning, as seen in the results for *roberta-base*, *distilroberta-base*, and the MultiBERT models.

3.3.3 Summary. It is clear that processing both the head and modifier in the same context tends to strengthen the correlation between the resulting compound representation and the broad semantic relation category RDM for most layers of most models. This "compositional gain" is strongest with BERT-style models, where the difference in correlation value was significant for most layers. This finding sheds light on the first part of our second research question, where we predicted that some models would contextually compose head/modifier semantic information rather than memorizing distributional co-occurrence information about head/modifier words and their associated semantic relations.

3.4 Experiment 2a: Relation Vector RSA

3.4.1 Overview. In this experiment we use RSA and the 60 compound dataset to measure the representation of the fine-grained relation vectors across different models, layers, choices of representation, and levels of granularity. To measure different levels of granularity, we target two ground-truth RDMs: (a) an RDM using the top-mentioned thematic relation dimension in the thematic relation vector for each compound (created by considering two compounds to be maximally similar if they share their most frequently reported relation, and maximally different otherwise) and (b) an RDM using the full 18-dimensional relation vector for each compound (created by measuring pairwise cosine similarity between compounds). We can consider the correlation between model-elicited RDMs and the top-mentioned relation RDM as a measure of how well the Transformer-based language models encode a more coarse-grained representation of noun-noun compound semantic similarity across a broad variety of compounds, thus acting as a bridge between the Relation Category experiments (Sections 3.2 and 3.3) and the fine-grained 18-dimensional RSA of the Relation Vector RSA experiments. The 60×60 RDMs can be seen in Figure 7. As in the Relation Category RSA experiment, the data for the experimental RDMs is calculated as the model activation patterns for the mean-pooled token spans across (1) the modifier word, (2) the head noun, and (3) the whole compound. We construct three experimental RDMs for each layer of each model by taking the cosine similarity between all 3,600 pairs of samples for the three choices of representation and use Pearson's r to correlate the experimental RDMs with the ground-truth RDMs. Again, we only consider the upper triangle (excluding the main diagonal) of each RDM in our correlations.

3.4.2 Results. The results for the Relation Vector RSA experiment are given in Figure 8. Overall, we see the same pattern of results for this dataset and these ground-truth RDMs as in the Relation Category RSA experiments: *roberta-base* and *distilroberta-base* show the overall strongest correlations, followed by the MultiBERTs models and *xlnet-base-cased*, and the poorest fit to the ground-truth RDMs is seen for the non-monolingual models. As in both parts of the Relation Category experiments, the baseline Japanese model (*bert-base-japanese*) does not provide strong correlations between the model activation RDMs and the ground-truth RDMs, indicating that this model fails to encode information about the kind of thematic relation used in compounds. Similarly, the multilingual transformer model (*xlm-mlm-xnli15-1024*) also achieves relatively low correlations in all layers when compared to the four models trained only on English corpora.

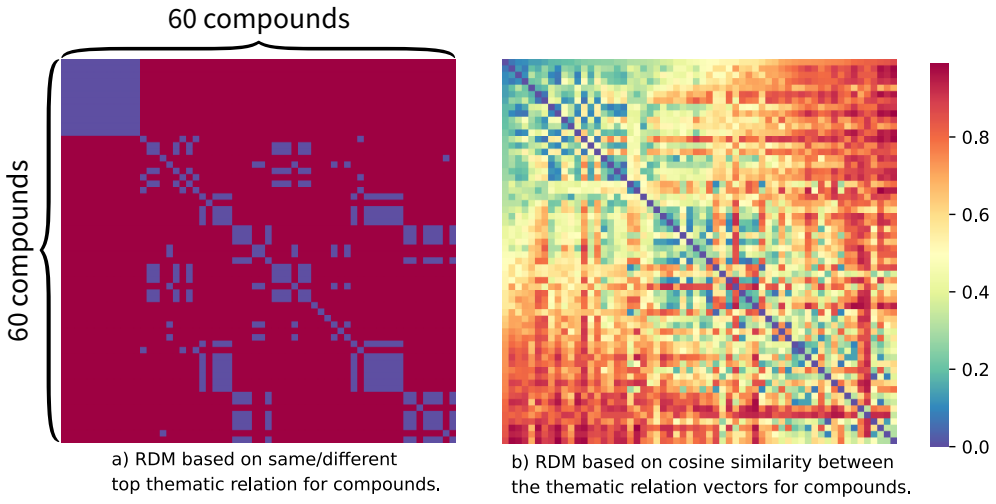


Figure 7
RDMs based on the "fine-grained" semantic relation vector representations, using 60 compounds.

Overall there are stronger correlations for model representations for the RDM based on the top-mentioned relations (shown by solid lines in Figure 8) than for the RDM based on full thematic relation vectors (shown in dashed lines). This is particularly apparent in BERT-style models and in early layers of *xlnet-base-cased* and *xlm-mlm-xnli15-1024*. One interesting trend in the results of the Relation Vector RSA is that representations from *xlnet-base-cased* and *xlm-mlm-xnli15-1024* strongly distinguish compounds by their top-mentioned relation in earlier layers of processing before this correlation declines in a step-wise manner across layers. We also find that the fine-grained 18-dimensional representation of compounds is more strongly distinguished in the later layers of these same models. When taken together, these effects appear to be a trade-off between the two thematic relation signals in *xlnet-base-cased* and *xlm-mlm-xnli15-1024*, with the more general relation classification being strongly apparent at the beginning of processing before gradually giving way to a more fine-grained view of head-modifier semantic relations. The *xlm-mlm-xnli15-1024* result is somewhat surprising as the Relation Category RSA showed that this model's coarse-grained semantic signal follows a strong positive monotonic trajectory across layers, although in both analyses the correlations for *xlm-mlm-xnli15-1024* are not strong. We also observe some differences in trends between the Relation Category RSA and the Relation Vector RSA experiments with *xlnet-base-cased* (for example, the correlations achieved by *xlnet-base-cased* are more stable across layers in the Relation Category RSA). In contrast, none of the BERT-style models feature any such apparent discrepancy (although there is a greater variation in correlation strength across representation types for these three models). All three types of BERT model tend to show the strength of both the general and fine-grained thematic relation signal varying in the same direction together over the course of layers.

3.4.3 Summary. We found that the strongest correlations across most models were with the version of the RDM that only considered the top-mentioned relation dimension, for the BERT and RoBERTa models. Interestingly, the most strongly correlated

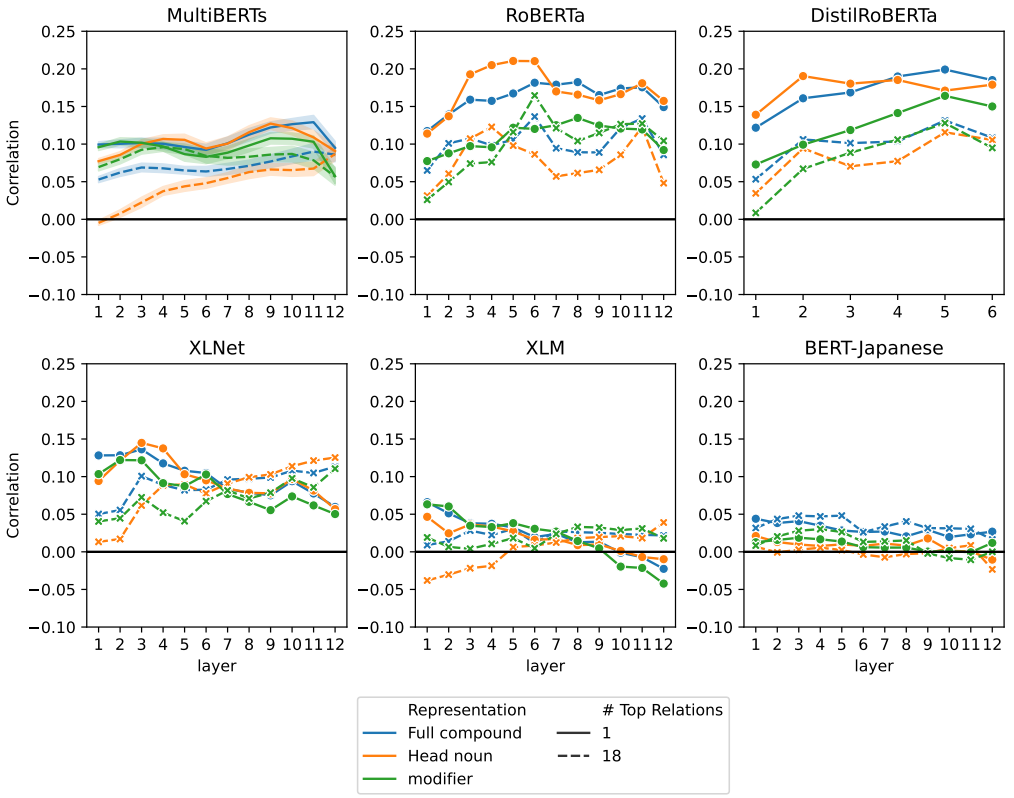


Figure 8
Results of the Relation Vector RSA experiment. Correlation between Transformer representation RDMs and the ground-truth semantic relation RDMs when (i) considering only the top mentioned thematic relation in each vector and (ii) similarity of the full 18-dimensional relation vectors.

representations of the broad semantic category were found in the head noun tokens of *roberta-base* and the representations that most aligned with the full 18-dimensional relation vector were found in the modifier nouns of that same model, suggesting that different types of relational information could be localized in different parts of the compound, a finding that is consistent across the BERT-style models.

3.5 Experiment 2b: Relation Vector and Processing Condition RSA

3.5.1 Overview. In this experiment we use RSA to measure the correlation between the 18-dimensional relation vectors and the Transformer model representations under the two processing conditions introduced in the Relation Category and Processing Condition RSA experiment (Section 3.3): (i) when the head and modifier nouns of a compound are processed in the same sentence, and (ii) when the head noun and modifier noun are processed in two separate sentences. In both processing conditions we take the mean-pooled intermediate token vector across head and modifier tokens as the compound representation. In this experiment we use the full 18-dimensional relation vector for each compound, as in condition (ii) of the previous Relation Vector RSA experiment.

3.5.2 Results. The results for the Relation Vector and Processing Condition RSA experiment are given in Figure 9. As in previous experiments, we find that *roberta-base* and *distilroberta-base* representations elicit the highest correlation strengths (when compound components are processed together in the same context), and that *bert-base-japanese* achieves relatively low correlations. In contrast to the relatively strong coarse-grained semantic signal in later layers of *xlnet-base-cased* that were observed in the Relation Category RSA experiments, we see that *xlnet-base-cased* struggles to represent the fine-grained semantic signal in both processing conditions, often resulting in lower correlations than the baseline Japanese monolingual model. As was seen in the Relation Category RSA experiments, representational dissimilarity patterns produced by *xlnet-base-cased* and the MultiBERT models tend to align more with the semantic relation RDM than the multilingual and monolingual Japanese model, but less than *roberta-base*.

By examining the effect of the processing condition on the representation of fine-grained semantic differences between compounds, we find that processing the head and modifier words in the same context almost always leads to a stronger semantic signal in the final compound representation. The trends observed in this fine-grained setting broadly align with the results of the Relation Category and Processing Condition experiment, where it was found that *roberta-base* and *distilroberta-base* benefit massively from the same-context processing condition. In the Relation Vector and Processing

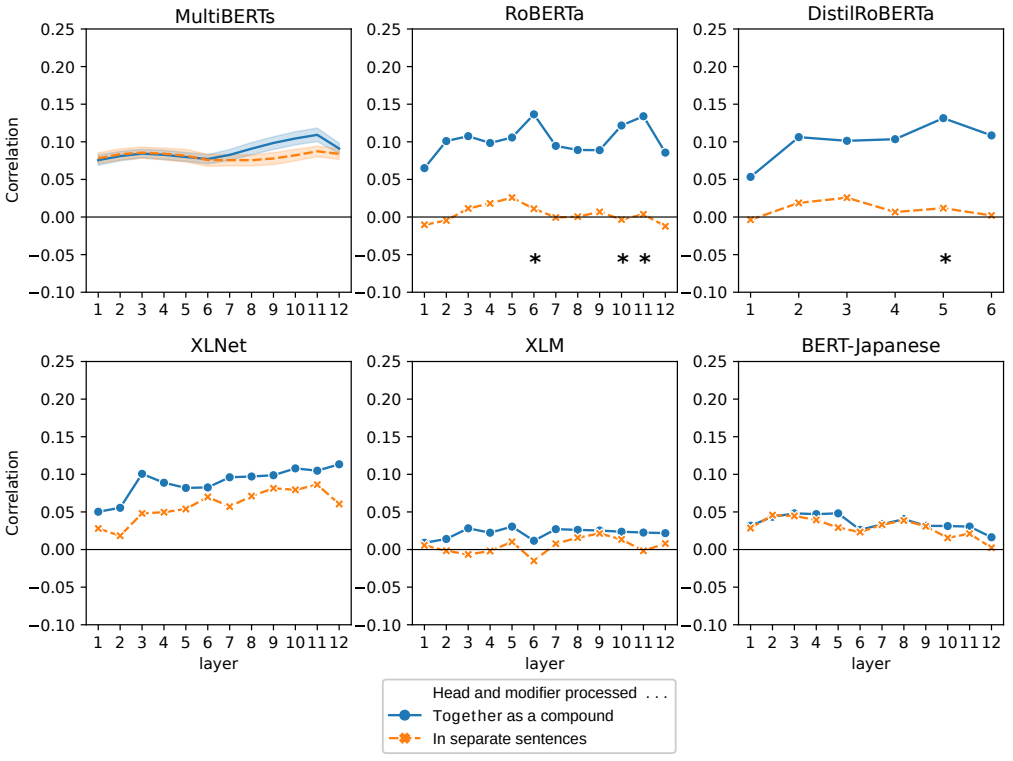


Figure 9
Results of Relation Vector and Processing Condition experiment. Correlation between Transformer representation RDMs and the ground-truth semantic relation RDM (using all 18 relations) under two processing conditions.

Condition experiment, we find fewer significant differences between the processing conditions than in the Relation Category version of the experiment, although all significant differences are found in *roberta-base* and *distilroberta-base*. While *xlnet-base-cased* and the MultiBERTs tend to improve their representation of both the coarse-grained and fine-grained semantic RDMs under the normal same-context processing condition, they tended to benefit less than the RoBERTa style models in the Relation Category and Processing Condition RSA. This effect is more pronounced in the fine-grained setting, where the MultiBERT models in particular produce relatively good representations of the 18-dimensional relation vectors when its head and modifier words are processed separately in different sentences before being mean-pooled together. In contrast, separate-context representations for both *roberta-base* and *distilroberta-base* never reach a correlation strength greater than 0.02. By comparing this result to the Relation Category and Processing Condition RSA, we can argue that RoBERTa-based models can (to an extent) encode broad semantic categories in singular word representations while failing to represent almost any of the fine-grained semantic relations that could apply to a given head or modifier word until a corresponding modifier or head-noun is provided in the same context. On the other hand, the MultiBERT models and *xlnet-base-cased* represent potential relations that can apply to a particular head or modifier in static representations, despite the corresponding modifier or head noun not being seen in the same processing context. One of the biggest differences between the results of the Relation Category and Relation Vector versions of the experiment is that *xlm-mlm-xnli15-1024* is able to produce relatively good representations of coarse-grained noun-compound semantic distinctions (particularly in later layers), while failing to capture much of fine-grained semantic differences between compounds.

3.5.3 Summary. As was found in the Relation Category RSA experiment (Section 3.3), correlations between the relation vector RDM and the model-elicited RDMs were generally stronger when the head and modifier were processed in the same context, again agreeing with our prediction that allowing the whole compound to be compositionally processed leads to the model producing representations that encode more information about the semantic category of the compound. This "compositional gain" was particularly strong in *roberta-base* and *distilroberta-base*, but surprisingly there was little drop-off in correlation strength when the MultiBERT models processed head and modifiers in different contexts.

3.6 Experiment 3: Compositional Probe

3.6.1 Overview. Using a complementary methodology to the RSA-based analyses of the previous experiments, we also conduct a Compositional Probe experiment that is designed to test whether mean-pooled token vectors corresponding to modifier words and head nouns require concurrent processing of both words in the same sentential context in order to encode fine-grained thematic relation information. To this end, a probing experiment is defined that uses linear regression models to predict the 18-dimensional thematic relation vector (from the Devereux and Costello [2005] dataset) from mean-pooled token vectors across compound spans under the two processing conditions defined in the Relation Category/Vector and Processing Condition RSA experiments: (1) when the head and modifier word are processed normally as a compound in the same sentence and (2) when the head and modifier word are processed in separate sentences before being mean-pooled.

Our methodology uses an adapted version of the 2 vs. 2 test framework described in Mitchell et al. (2008) and Xu, Murphy, and Fyshe (2016). For a given set of compound

representations of size $n = 60$ we carry out linear regression probing tests that compare all possible pairs of compounds (1,770 pairs in total). For each unique pair consisting of compound i and compound j, we train a linear regression model to predict the relation vectors for the other 58 compounds using the corresponding 58 compound representations. The model then produces predictions $\tilde{Y}^i$ and $\tilde{Y}^j$ from X^i and X^j, and we evaluate whether a test is successful based on the criterion:

$$dist(\tilde{Y}^i, Y^i) + dist(\tilde{Y}^j, Y^j) \quad < \quad dist(\tilde{Y}^i, Y^j) + dist(\tilde{Y}^j, Y^i)$$

where the distance is measured using mean-squared error. Note that when the success criterion is met, the regression model produces relation vector predictions for compounds i and j that are closer to the true relation vectors for compounds i and j, respectively, than the other way round.

We run this set of probing tests for each layer of each model and compare the decodability of the normally processed compound representations to the compound representations processed over two separate sentences. If both types of compound representations achieve similar numbers of successful tests in this experiment, this would indicate that the representations of the head noun and modifier word separately encode the range of common thematic relation types for each word, and that this encoding does not depend on the compositional meaning of the two words together in a compound. For example, the word MOUNTAIN as a modifier may tend to often be used with a *M located in H* relation in compounds (as in the phrases MOUNTAIN STREAM, MOUNTAIN CABIN, etc.), and the models may be sensitive to this kind of thematic information in their representation of individual words, without representing the relation in specific compounds. On the other hand, if it is much easier to decode the relation in contextually processed noun-noun compounds, then we would argue that the model instead encodes thematic relation information using a contextually aware composition mode.

3.6.2 Results. The results for the Compositional Probe experiment are given in Figure 10. For these results, for each model and layer, we statistically test whether the number of successful 2 vs. 2 tests (from 1,770 tests in total) for the condition where the modifier and head are presented together as a compound is greater than the number of successes when the modifier and head are processed in separate sentences. In this statistical analysis, there are dependencies in the outcomes of the 1,770 tests for the two conditions that need to be taken into account. Firstly, the outcomes for the two conditions are *paired*; the outcome for a test for a particular pair of compounds (i, j) in the Together condition is not independent of the corresponding outcome in the Separate condition, as they involve the same lexical items. Secondly, the probability of a success for a given pair of compounds (i, j) will depend on the quality of the language model's representation of compounds i and j, and this will vary from compound to compound. A consequence of this is that outcomes are not statistically independent across the 1,770 tests (for example, if a language model has a poor representation for compound i, then this means that the probability of a success for the 59 tests containing compound i will be low, compared with tests not containing this compound).

In order to take these statistical dependencies into account, we perform a randomization test (Edgington and Onghena 2007) to compare the number of successes across the two conditions. Our null hypothesis is that the number of successes in the 'Together' and 'Separate' conditions do not differ. Under this null hypothesis, the probability of a success in the two conditions *for a given pair* does not differ, and thus the observed

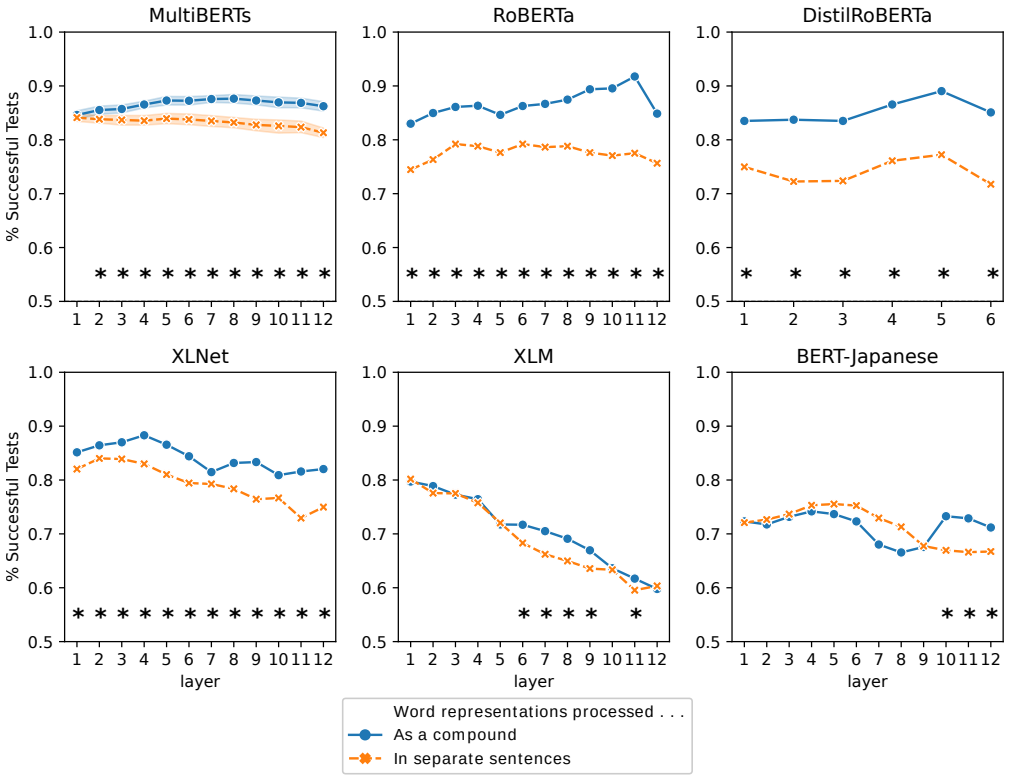


Figure 10
Results of the Compositional Probe experiment. The proportion of successes across 1770 2-vs-2 tests to decode thematic relation vectors from compound representations. Results are reported for two compound processing conditions: (i) when the head noun and modifier noun are processed together as a single compound in the same sentence, and (ii) when the head noun and modifier noun are processed in two separate sentences. Asterisks mark significant differences (at the threshold of $p < 0.05$) between the number of successes across the two processing conditions. Chance performance on this experiment is 0.5.

outcome for that pair in the Together condition is interchangeable with the observed outcome for that pair in the Separate condition. Thus, in one run of our randomization procedure, the two outcomes for each pair are randomly assigned to the two conditions, and we calculate the difference between the number of successes in the two conditions. We perform 10,000 runs of this randomization procedure, to build a distribution of these differences assuming the null hypothesis is true. Finally, we obtain a p-value for a one-sided test testing whether the actual observed number of successes for the Together condition is greater than the Separate condition, by counting the proportion of times the observed difference is greater than the differences obtained across our 10,000 randomization runs. As in the earlier experiments, significant effects at $p \leq 0.05$ after applying a false discovery rate controlling procedure (Benjamini–Hochberg with $\alpha = 0.05$) are indicated with asterisks in Figure 10.

For all layers of the four monolingual English Transformers, the thematic relation vector is more decodable from the compound representation when the head and modifier words are processed together in the same sentence. This compositional gain from

same-context processing was found to be statistically significant in all layers of these four models excluding the first layer of the MultiBERT models. Within the four most decodable models, representations from the MultiBERT models gain the least from the normal contextual processing condition. Despite the relatively low difference in successful tests for the MultiBERT models across the two conditions, the fact that these differences are significant across the 25 instantiations of the same model suggests that this effect is robust and that BERT models consistently produce better representations of the semantic relation vector when the head and modifier words are processed together as a compound. Alongside the RSA analyses of previous experiments, these results again indicate that the Transformer-based language models that most strongly encode compound semantic relations also tend to compositionally integrate their knowledge of the head noun and modifier word in order to represent semantic relation information, above and beyond what can be decoded by relying on any association between thematic relations and the individual words.

Although the results above demonstrate better than chance decoding accuracy in the 2 vs. 2 experiment for most models and layers, we next investigated which individual compounds had poor quality predicted relation vectors, to better understand how the Transformer-based language models may fail to capture the compositional semantics of compounds. For the 2 vs. 2 decoding experiment, we save the predicted vector $\tilde{Y}^i$ of every compound i, averaging the predicted vectors across the 59 tests where compound i appears in the pair of compounds. In this way, we obtain an average relation vector prediction for each of the 60 compounds for each layer of each language model. In an exploratory analysis to investigate which compounds had generally poor quality predicted relation vectors across models and layers, we used the DBSCAN algorithm (Ester et al. 1996) to perform 60 cluster analyses—for each compound, we cluster all the predicted relation vectors, for every type of model and layer. For this analysis we select one candidate *bert-base-uncased* model from the set of MultiBERTs (as opposed to averaging the models' predictions, which would be akin to constructing an ensemble model that would perform better than any particular BERT model). We next calculated the average predicted relation vector for each cluster to obtain cluster centroids, and ranked the compounds by the greatest amount of error incurred by the best-performing cluster (i.e., the cluster with the smallest Euclidean distance between the cluster centroid and the ground-truth relation vector).

The five compounds that were most difficult to decode in the compound decoding experiment from the Compositional Probe experiment are presented in Table 2. For the compound CONSTRUCTION EQUIPMENT the ground-truth relation vector has high values for the *H for M*, *H used by M*, and *H causes M* dimensions (i.e., EQUIPMENT *for* CONSTRUCTION, EQUIPMENT *used by* CONSTRUCTION and EQUIPMENT *causes* CONSTRUCTION). However, in the clustered prediction relation vectors, the closest cluster centroid has high values for the *H derived from M*, *H made of M*, and *H is M* dimensions. Comparing the cluster's prediction for the compound CONSTRUCTION EQUIPMENT to the ground truth relation vector of STEEL EQUIPMENT, we see that the models are likely to have been overfitted for the word EQUIPMENT (i.e., equipment tends to be made of something else, but this is not actually reflected in the semantic relationship with *construction*). Similarly, the predictions for STEEL EQUIPMENT seem to be informed by the ground truth relation vector of CONSTRUCTION EQUIPMENT. We also found a similar effect for VAPOR CLOUD and VAPOR DROPS. The other two difficult compounds, CREAM CHURN and BREAKFAST SUGAR, feature unique head and modifier nouns within the dataset, and as such have not been subject to this lexical bias. For the compound CREAM CHURN, the closest cluster contains predicted relation vectors from layers 3-8 of the

Table 2
Top five most difficult compounds (as measured by the distance between predicted and actual relation vectors) on the 60 compound linear regression relation vector decoding experiment (Section 10). Compounds are ranked by greatest average distance between the best-performing set of models (grouped by clustering their predictions). The top three relation dimensions and their values are reported for the ground truth and predicted relation vectors. The models in the best-performing cluster are abbreviated. Model layer ranges are given in superscript.

Compound	Ground truth	Cluster prediction	Models in cluster
construction equipment	*H FOR M (15)* *H USED BY M* (11) *H CAUSES M (7)*	*H DERIVED FROM M* (11.4) *H MADE OF M* (9.9) *H IS M* (8.2)	BBJ^{1-12} BBU^{1-12} DB^{1-6} RB^{1-12} XBC^{1-12} $XMX1^{1-12}$
steel equipment	*H MADE OF M* (15) *H DERIVED FROM M* (14) *H IS M* (11)	*H USES M* (8.9) *H FOR M* (5.1) *H USED BY M* (5.1)	BBJ^{1-11}
vapor cloud	*H MADE OF M* (17) *H IS M* (11) *H DERIVED FROM M* (8)	*H CAUSES M* (7.8) *H MAKES M* (7.5) *H USES M* (5.1)	BBJ^{1-12} BBU^{1-12} DB^{1-6} RB^{1-12} XBC^{1-12} $XMX1^{1-8}$
cream churn	*H MAKES M* (16) *H FOR M* (10) *H USES M* (5)	*H USES M* (10.7) *M CAUSES H* (6.3) *H HAS M* (4.5)	BBJ^{3-8}
breakfast sugar	*H FOR M* (12) *H DURING M* (10) *H USED BY M* (6)	*H DERIVED FROM M* (9.2) *M MAKES H* (8.3) *M CAUSES H* (8.1)	BBU^{1-12} DB^{1-6} RB^{1-12} XBC^{1-12} $XMX1^{1-10}$

bert-base-japanese model, indicating that the English language models are producing poor representations of the relation in this compound. While CHURN *makes* CREAM is the top dimension in the ground truth, CHURN *uses* CREAM is the top predicted dimension; this may reflect the relatively few total mentions for this *H makes M* relation type across the 60 compounds (as can be seen in Figure 2). The top predictions for BREAKFAST SUGAR all erroneously indicate that sugar is derived from breakfast. One possible explanation for such predictions is that relations such as *H derived from M*, *M makes H*, and *M causes H* are commonly associated with food concepts in the dataset (e.g., OLIVE PASTE, VEGETABLE APPETIZER & GRAIN CONTROVERSY). The pattern of errors underlines the importance of controlling for associations between individual words and semantic relations when probing the models for compositional semantics, as we do in our compositional probing technique and in our contrast of the Together and Separate processing conditions in the RSA analysis.

3.6.3 Summary. All four of the monolingual models produce representations of compound nouns that are more easily decoded for head-modifier relational information when the head and modifier words are processed in the same context. This effect is stronger than we anticipated given previous work on compositionality of multi-word expressions in Transformer models. This relational information is less available for probing in the baseline Japanese and the multilingual model, which often produce

representations that are just as decodable (and sometimes less decodable) when head and modifier words are processed separately rather than as a compound.

4. Principal Findings and Implications

Using a novel approach based on ground-truth human annotations of relation meaning in the interpretation of noun-noun compounds, we used complementary representational similarity and linear probing methods to investigate whether Transformer-based language models represent the semantics of the thematic relation in noun-noun compounds. Across the experiments and analyses, we find that the English-language Transformer models, and in particular *roberta-base* and *distilroberta-base*, consistently and significantly represent compound relation semantics. Importantly, this finding persists even with careful control of the lexical content of the compounds, achieved both through psycholinguistic design that orthogonalizes relational content and the modifier and head words (i.e., the Relation Category RSA experiments) and through comparing the quality of the compound relation representation to the non-compositional representations obtained when processing the modifier and head words separately (the Relation Category/Vector and Processing Condition experiments and the Compositional Probe).

4.1 Knowledge of Implicit Intra-compound Semantic Relations

Using the 300 compound dataset and representational similarity analysis, the Relation Category RSA experiments showed that models apart from *bert-base-japanese* produce representations that moderately correlate with simple coarse-grained thematic relation signals. In the Relation Vector RSA experiments, an alternate coarse-grained RDM was constructed by considering pairs of compounds from the 60 compound dataset to be similar only if they share their top-mentioned relation. Again, this experiment provided evidence that all five Transformer-based language models that were exposed to significant English language training data (i.e., all but *bert-base-japanese*) produce token vector representations of compounds that can be distinguished by the dissimilarity pattern induced by the top-mentioned ground-truth relation (although the multilingual model, *xlm-mlm-xnli15-1024*, achieved relatively low correlations with the thematic relation signal on this task compared to the four monolingual English models).

We also found evidence that Transformer-based language models learn multi-dimensional fine-grained aspects of the semantic relation between head nouns and modifier nouns in English noun-noun compounds using the 60 compounds relation vector dataset (Devereux and Costello 2005) coupled with RSA (in the Relation Vector RSA experiments) and linear regression probing models (in the Compositional Probe experiment). In the Relation Vector RSA experiments it was found that the 18-dimensional relation vector representation task was more difficult than distinguishing compounds by whether they share their primary relation. Despite this increased difficulty, it was shown that token vector representations from four of the monolingual English models achieve moderate correlations with the 18-dimensional representational dissimilarity matrix. While *xlm-mlm-xnli15-1024* only achieved correlations of around one-third the strength of the most highly correlated model (*roberta-base*), this model demonstrated evidence of increasingly stronger correlations towards later layers. On the other hand, the baseline non-English monolingual model, *bert-base-japanese*, achieved a consistently low correlation strength, which may have been inflated slightly by a lexical overlap bias. In the Compositional Probe experiment, we found clear evidence that every model

apart from the Japanese model produces token vectors that can be used to predict the 18-dimensional compound with linear regression probing models. Again, in this task the BERT-style models excelled over *xlnet-base-cased* and *xlm-mlm-xnli15-1024*. Taken together, these results clearly show that Transformer-based language models learn to encode information about the semantic relation between head nouns and modifier words in English noun-noun compounds. This finding conflicts (to an extent) with Yu and Ettinger (2020), where little evidence of compositionality was found in Transformer model representations using similarity ratings and paraphrase classifications. In contrast to that work, the present analysis uses an explicit model of compound semantics based on human annotation data for thematic relations, allowing us to directly measure the semantic representation for a given compound. However, in cases where Yu and Ettinger (2020) identify limited evidence of compositionality, they find that RoBERTa outperforms both XLM and XLNet, a result that generally aligns with our findings.

4.2 Encoding Mechanisms for Intra-compound Semantic Information

Our second major research question was how information about the semantic relation between the head and modifier word in a compound noun was encoded in Transformer-based language models. We identified three main areas of interest within this investigation: (1) whether the representation of this semantic relation results from a dynamic composition mode rather than relying on memorizing distributional co-occurrence information, (2) if this information is primarily localized within a particular token span within the compound vectors (i.e., in the head or modifier token vectors), and (3) whether this information was generally localized to a particular layer or set of layers within Transformer-based language models.

We investigated the question of whether Transformers dynamically compose information about the thematic relation between head words and modifier words by developing two types of "compositional probes" that check for statistical differences between how well a Transformer represents this semantic information under two processing conditions: (1) when a head and modifier words are processed normally as a single compound in the same context, and (2) when the head and modifier words are processed in separate sentences before their token vectors are mean-pooled. We argue that if a Transformer-based language model encodes more relational information about the compound under the same-context processing condition, then this model does not rely solely on distributional information about the co-occurrence of particular head/modifier words and their likelihood to be used with particular semantic relations; instead, they must be representing compositional relational information that is true of the compound as a phrase. In the RSA version of this compositional probe (the Relation Category and Processing Condition RSA experiment), we found that almost all layers of the English monolingual models benefited significantly from same-context processing condition. In contrast, only the final few layers of the multilingual model showed a compositional gain (a difference which was not statistically significant) and the baseline Japanese model achieved around chance levels of decodability under both conditions. Of the models that did demonstrate a significant compositional gain in our compositional probe (i.e., the normal compound processing condition), *roberta-base* and *distilroberta-base* both demonstrated the largest compositional gain and the overall best correlation to the ground-truth relation dissimilarity matrix, although the MultiBERTs models and *xlnet-base-cased* were relatively easily decodable despite benefiting less from processing constituent words of a compound in the same sentence. In the linear

regression decoding version of this analysis (the Compositional Probe experiment), we found evidence that most layers of all models except *xlm-mlm-xnli15-1024* and *bert-base-japanese* benefit significantly from dynamic compositional processing in the 60-compound setting. These results generally show that Transformer-based language models produce representations of compounds that best encode for semantic information about how the head word relates to the modifier word when these constituent words are processed in the same context.

Another question of interest within the overall enquiry into how these models encode semantic information relating to head and modifier words is to what extent this information is localized to particular token spans within a Transformer's representation of a noun-noun compound. In order to shed light on this area, we included representations from head and modifier word token spans in our Relation Category and Relation Vector experiments. In these experiments we investigate the representation of both the coarse-grained and fine-grained semantic relation signal using our two main analysis techniques (RSA and linear decoding). In the Relation Category RSA experiment, it is difficult to point to any broad representational trends other than that the correlation strengths of the modifier noun vector and the full compound vector were similar in most of the Transformer-based language models we investigated. Nonetheless, the fact that the baseline Japanese model achieves a relatively high correlation with the relation type RDM in the Relation Category RSA experiment is somewhat conspicuous. In the Relation Vector RSA experiments, we observe a pattern across the BERT-style models whereby the head noun is preferred for representing the shallow top-mentioned relation RDM, whereas the modifier noun in many cases elicits stronger correlations than even the whole-compound vector. These trends however do not tend to hold for *xlm-mlm-xnli15-1024* and *xlnet-base-cased*, where the differences are more difficult to interpret.

The final aspect of how implicit intra-compound thematic relation information is encoded that we investigated is whether this information is localized to particular layers of processing. Across our experiments we find disparate trends in the results with respect to both correlation with ground-truth RDMs and decodability scores across layers. Among the BERT-style models, we find that various measures of the semantic relation signal are strongest in early/middle layers of the MultiBERT models and middle/late layers of *roberta-base* and *distilroberta-base*. In the Relation Category RSA experiment, we can see that the coarse-grained thematic relation signal of the MultiBERT models is at its strongest in the middle layers, before this information diminishes in the final few layers of processing. This result appears to contrast with Tenney, Das, and Pavlick (2019), who found that semantic information appears in later layers of BERT. We note however that performance on the semantic tasks used in that work (i.e., semantic role labeling [SRL] and coreference) reflect a model's ability to process high-level semantic information. It could be the case that a model's capacity to distinguish between words using semantic concepts such as thematic relation is a prerequisite to capturing the high-level semantic information required in SRL and coreference resolution. Furthermore, Tenney, Das, and Pavlick (2019) note that semantic information is dispersed widely across layers (compared to the stronger localization of information associated with syntactic processing). This phenomenon can be seen in the results of the four English-language monolingual models in the Relation Category RSA experiment, where we can recover a relatively good representation of the coarse-grained thematic relation signal from the least correlated layers. Interestingly, we find several cases where the layer-wise correlation trends of *xlnet-base-cased* align with those seen in *xlm-mlm-xnli15-1024*. In particular, both of these models show a coarse-grained/fine-grain trade-off in the Relation Vector RSA experiments, and both of these models decrease in decodability

across subsequent layers in an almost monotonic fashion in the Compositional Probe experiment. In general, both *xlm-mlm-xnli15-1024* and *xlnet-base-cased* tend to represent fine-grained dissimilarity patterns (as measured with RSA) better in later layers. At the same time, the individual dimensions of vectors produced by these two models become less decodable in later layers. In the case of the BERT-style models, it is clear that the final layer is not the best for representing the thematic relation signal and we recommend exploring the use of middle layers for downstream tasks that require similar relational-semantic information.

5. Limitations and Future Work

One of the main limitations of our analysis was the relatively small size of the datasets we used compared to other datasets that are used for evaluation in natural language processing research. In this area we are limited by the lack of large, annotated noun-noun compound datasets, as the process of labeling noun-noun compounds with thematic relations is a time-consuming process for human annotators. This is particularly the case for the 18-dimensional setting, where every potentially applicable relation must be considered, rather than choosing one main relation. The fact that we can only use the 300 compound dataset to compare within each of the 60 groups means that we are limited to a total of 540 comparisons between compound representations using RSA, as the upper triangle of the ground truth RDM structure in Figure 4 (excluding the main diagonal and the compound pair marked in gray) allows for 9 comparisons within each group. We identify this annotation task as a key recommendation for future work for extending our analyses. While the 60 compound dataset provides a richer annotation of the underlying relation between head and modifier nouns, this dataset again allows for relatively few comparisons (1,770 pairs for RSA and the 2 vs. 2 test). In part due to the limited number of samples available to us, we chose to use data analysis techniques developed by researchers in the area of cognitive science, as this field is often limited by both the number of subjects that can be observed, and the number of stimuli that can be presented to a human subject within a single session.

Another area where our analysis is limited is the range of Transformer-based language models we investigated. While these models were chosen to cover a range of Transformer types, it is impossible to make generalizable judgments about certain classes of model (i.e., multilingual models or distilled models) based on our analysis, as we only feature one type of model in each of those classes. One exception is our analysis of *bert-base-uncased*, *roberta-base*, and *distilroberta-base*, which allows us to make generalizable statements about this class of Transformer. This is particularly true of the *bert-base-uncased* model, as we carry out an analysis on 25 different instantiations of this same class of model. In any case, future work should expand the analysis of how Transformers process noun-noun compounds to cover several models within each one of these areas. A related recommendation towards building a more robust analysis is to train several versions of the other five types of model, allowing for variation within a constrained architecture and choice of hyperparameters (McCoy, Min, and Linzen 2020).

As part of our experimental design, we do not consider the effect of fine-tuning Transformer-based language models. This choice allows us to probe Transformer-based language models for their capacity to automatically capture implicit semantic information about noun-noun compounds at the expense of limiting the generalizability of our findings in fine-tuning settings that are commonplace in the application of Transformer-based language models for solving downstream tasks. This consideration is particularly

relevant in the context of probing representations with simple linear models, as latent high-level semantic information about noun-noun compounds may require non-linear processing facilitated by fine-tuning several layers in order for this information to become available for linear regression models in the final layers. After juxtaposing the high RSA correlations found in later layers of *xlm-mlm-xnli15-1024* and *xlnet-base-cased* in the Relation Vector RSA experiments with the relatively low decodability scores for later layers of these two models seen in the Compositional Probe experiment, we would expect fine-tuning to be particularly useful for these models. One further consideration for expanding the analysis to allow for fine-tuning is that any sensitivity to noise incurred by a small amount of training samples may be amplified during this fine-tuning process, and thus the thematic relation vector dataset may need to be expanded before introducing this experimental extension. Another potential limitation with our experimental design is the choice to limit our RSA and probing tasks to token spans within the noun-noun compounds. As was seen in all experiments, Transformer-based language models tend to compose and distribute semantic information across many token vectors. Accordingly, it could be the case that information about the thematic relation between the head and modifier word could be distributed across tokens in other parts of the sentence, rather than just in tokens in the compound. A related concern to our choice of representation is the question of whether mean-pooling across tokens preserves thematic relation information and whether other approaches should be explored, such as concatenating a max-pooled vector with the mean-pooled token, or constructing a non-linear recurrent neural network probe to preserve all token vector information. In our experiments, we chose to consider a priori one choice of token representation that is most simply and most straightforwardly related to the compound (i.e., mean pooled token representations from within the compound), in order to avoid complications due to "researcher degrees of freedom" (Wicherts et al. 2016). Nevertheless, future work could investigate a wider range of possible model representations, to examine the extent to which these Transformer-based language models distribute information about compound relation semantics across the sentence, and whether such information is recoverable from a whole-sentence representation.

6. Conclusion

In this work, we used two English noun-noun compound datasets in order to probe Transformer-based language models for their knowledge of semantic relations between head and modifier nouns. To this end, we constructed three experiments to measure the representation of semantic relation information at a coarse and fine-grained level. In our layer-wise analysis, we find evidence that head-modifier thematic relation information is encoded in the token vector representations of six different Transformer-based language models. Of the six models we looked at, we find that the four English monolingual models strongly represent this information at both the coarse and fine-grained levels. Our compositional probe experiment shows that representations of these four models significantly benefit from head and modifier nouns being processed in the same context on a relation vector decoding task. Furthermore, we find evidence that these models gain significant levels of decodability from this concurrent compositional mode. These results suggest that the models that best encode relational information dynamically integrate their knowledge of the intrinsic properties of the head and modifier concepts in order to represent the semantic relation between these words, rather than only relying on distributional information of concept-relation frequency.

References

Abnar, Samira, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203. https://doi.org/10.18653/v1/W19-4820

Alishahi, Afra, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad. 2020. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Anderson, Andrew James, Douwe Kiela, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Edmund C. Lalor. 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, 41(18):4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021, PubMed: 33753548

Baroni, Marco. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307. https://doi.org/10.1098/rstb.2019.0307, PubMed: 31840578

Coil, Jordan and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do Language models understand noun compounds? *arXiv preprint arXiv:2305.10568*. https://doi.org/10.18653/v1/2023.findings-acl.169

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. https://doi.org/10.18653/v1/2021.emnlp-main.49

Devereux, Barry and Fintan Costello. 2005. Investigating the relations used in conceptual combination. *Artificial Intelligence Review*, 24(3–4):489–515. https://doi.org/10.1007/s10462-005-9007-5

Devereux, Barry and Fintan Costello. 2006. Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. In *28th Annual Conference of the Cognitive Science Society*.

Devereux, Barry J. and Fintan J. Costello. 2012. Learning to interpret novel noun-noun compounds: Evidence from category learning experiments. In *Cognitive Aspects of Computational Language Acquisition*. Springer, pages 199–234. https://doi.org/10.1007/978-3-642-31863-4_8

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language*, pages 810–842. https://doi.org/10.2307/412913

Edgington, Eugene and Patrick Onghena. 2007. *Randomization Tests*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420011814

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *The International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231.

Estes, Zachary and Uri Hasson. 2004. The importance of being nonalignable: A critical test of the structural alignment theory of similarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1082–1092. https://doi.org/10.1037/0278-7393.30.5.1082, PubMed: 15355137

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

Fares, Murhaf, Stephan Oepen, and Erik Velldal. 2018. Transfer and multi-task learning for noun–noun compound interpretation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1488–1498. https://doi.org/10.18653/v1/D18-1178

Gagné, Christina L. and E. J. Shoben. 1997. Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:71–78. https://doi.org/10.1037//0278-7393.23.1.71

Gagné, Christina L. 2001. Relation and lexical priming during the interpretation of noun–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):236. https://doi.org/10.1037//0278-7393.27.1.236, PubMed: 11204100

Gauthier, Jon and Roger Levy. 2019. Linking artificial and human neural representations of language. 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 529–539. https://doi.org/10.18653/v1/D19-1050

Girju, Roxana, Ana-Maria Giuglea, Marian Olteanu, Ovidiu Fortu, Orest Bolohan, and Dan Moldovan. 2004. Support vector machines applied to the classification of semantic relations in nominalized noun phrases. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 68–75. https://doi.org/10.3115/1596431.1596441

Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4. https://doi.org/10.3389/neuro.06.004.2008, PubMed: 19104670

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lees, Robert B. 1960. The grammar of English nominalizations. *International Journal of American Linguistics*, 26(3):205.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Li, Siyan, Riley Carlson, and Christopher Potts. 2022. Systematicity in GPT-3's interpretation of novel English noun compounds. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 717–728. https://doi.org/10.18653/v1/2022.findings-emnlp.50

Linzen, T. A. L. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):e99–e108. https://doi.org/10.1353/lan.2019.0015

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692 [cs]*.

Lynott, Dermot and Louise Connell. 2010. Embodied conceptual combination. *Frontiers in Psychology*, 1:212. https://doi.org/10.3389/fpsyg.2010.00212, PubMed: 21833267

Maguire, Phil, Barry Devereux, Fintan J. Costello, and Arthur Cater. 2007. A re-analysis of the CARIN theory of conceptual combination. *Journal of Experimental Psychology, Learning Memory and Cognition*, 33:811–821. https://doi.org/10.1037/0278-7393.33.4.811, PubMed: 17576155

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. https://doi.org/10.18653/v1/2020.blackboxnlp-1.21

Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Wikitext-103. Technical report, Salesforce.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x, PubMed: 21564253

Mitchell, Tom M., Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of Nouns. *Science*, 320(5880):1191–1195. https://doi.org/10.1126/science.1152876, PubMed: 18511683

Murphy, Gregory L. 1988. Comprehending complex concepts. *Cognitive Science*, 12(4):529–562. https://doi.org/10.1016/0364-0213(88)90012-2

Murphy, Gregory L. 2002. *The Big Book of Concepts*. MIT Press, Boston, MA. https://doi.org/10.7551/mitpress/1602.001.0001

Murty, Shikhar, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2022. Characterizing intrinsic compositionality in transformers with tree projections. *arXiv preprint arXiv:2211.01288*.

Nakov, Preslav. 2019. Paraphrasing verbs for noun compound interpretation. *arXiv preprint arXiv:1911.08762*.

Nili, Hamed, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553. https://doi.org/10.1371/journal.pcbi.1003553, PubMed: 24743308

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. https://doi.org/10.18653/v1/2022.acl-long.251

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Reddy, Siva, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sellam, Thibault, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, and Dipanjan Das. 2021. The multiBERTs: BERT reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Shwartz, Vered and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211. https://doi.org/10.18653/v1/P18-1111

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

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Tratz, Stephen and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687.

van Jaarsveld, Henk J. and Gilbert E. Rattink. 1988. Frequency effects in the processing of lexicalized and novel nominal compounds. *Journal of Psycholinguistic Research*, 17(6):447–473. https://doi.org/10.1007/BF01067911

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances In Neural Information Processing Systems*.

Westerlund, Masha and Liina Pylkkänen. 2017. How does the left anterior temporal lobe contribute to conceptual combination? Interdisciplinary perspectives. In *Compositionality and Concepts in Linguistics and Psychology*. Springer, Cham, pages 269–290. https://doi.org/10.1007/978-3-319-45977-6_11

Wicherts, Jelte M., Coosje L. M. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel A. L. M. Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, page 1832. https://doi.org/10.3389/fpsyg.2016.01832, PubMed: 27933012

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu, Haoyan, Brian Murphy, and Alona Fyshe. 2016. BrainBench: A brain-image

test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021. https://doi.org/10.18653/v1/D16-1213

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

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. https://doi.org/10.18653/v1/2020.emnlp-main.397

Yu, Lang and Allyson Ettinger. 2021. On the interplay between fine-tuning and composition in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2279–2293. https://doi.org/10.18653/v1/2021.findings-acl.201

Ó Séaghdha, Diarmuid. and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 57–64. https://doi.org/10.3115/1613704.1613712