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.

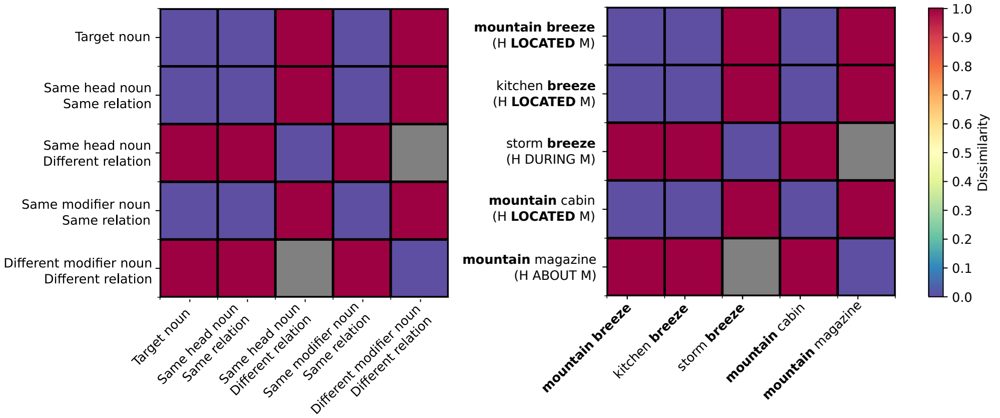
3.1 Experimental Design and Controls

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

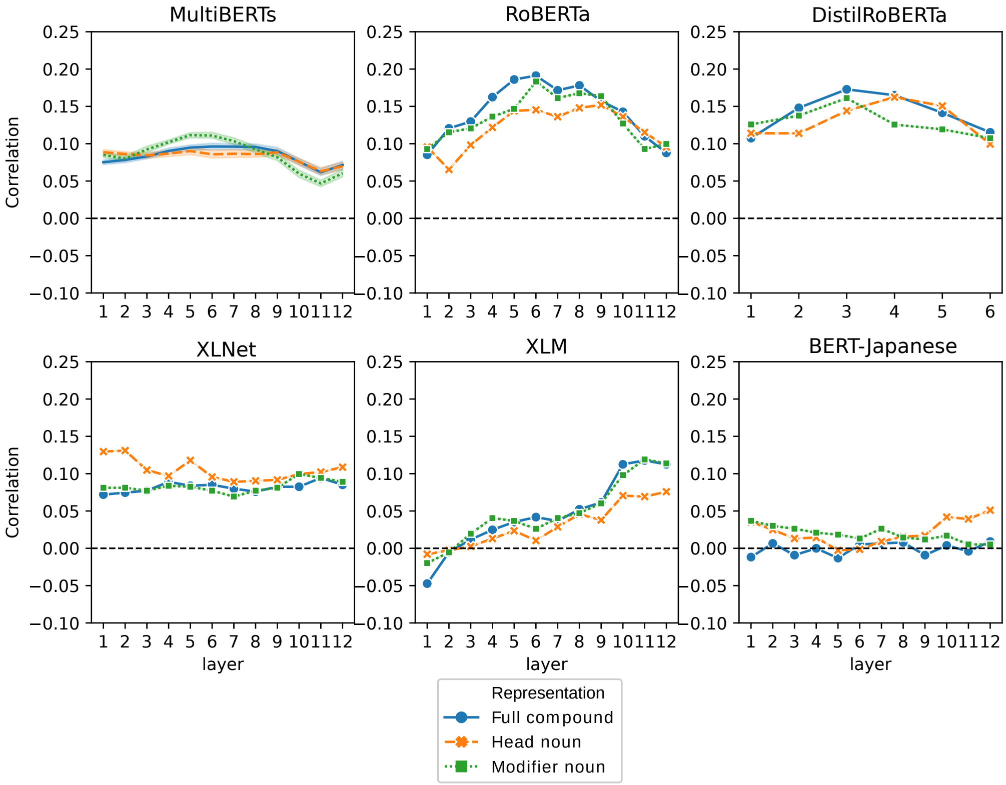


Figure 5
Results of the Relation Category RSA experiment (Section 3.2). Average correlation between the same thematic relation ground-truth RDM and experimental RDMs constructed using mean-pooled token-span representations for 6 types of Transformer-based language models (300 sentences, correlation averaged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