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.

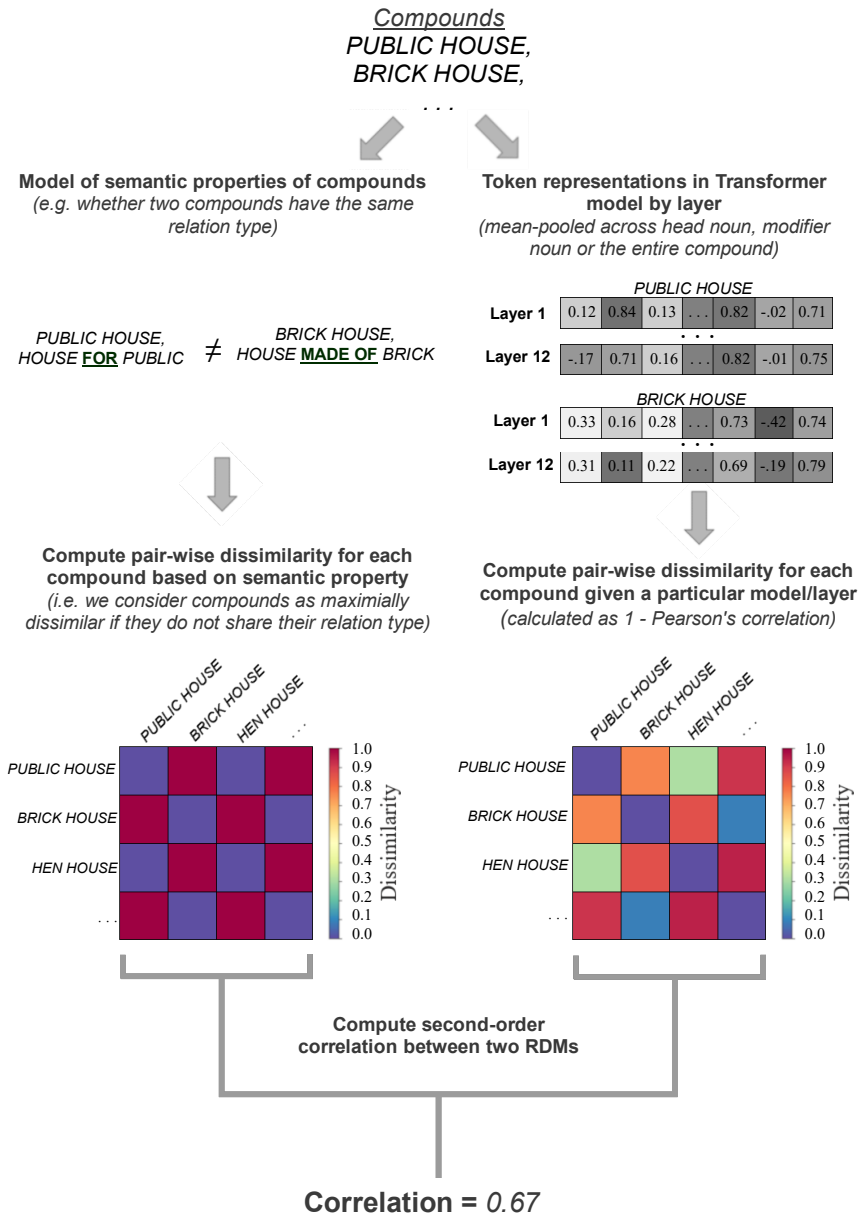


**Figure 1**
Overview of our feature extraction and Representational Similarity Analysis pipeline. Here we show the procedure for calculating the RDM for the Relation Category experiments (Experiments 1a and 1b), where compounds are counted as maximally dissimilar if they do not share a primary thematic 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).