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.

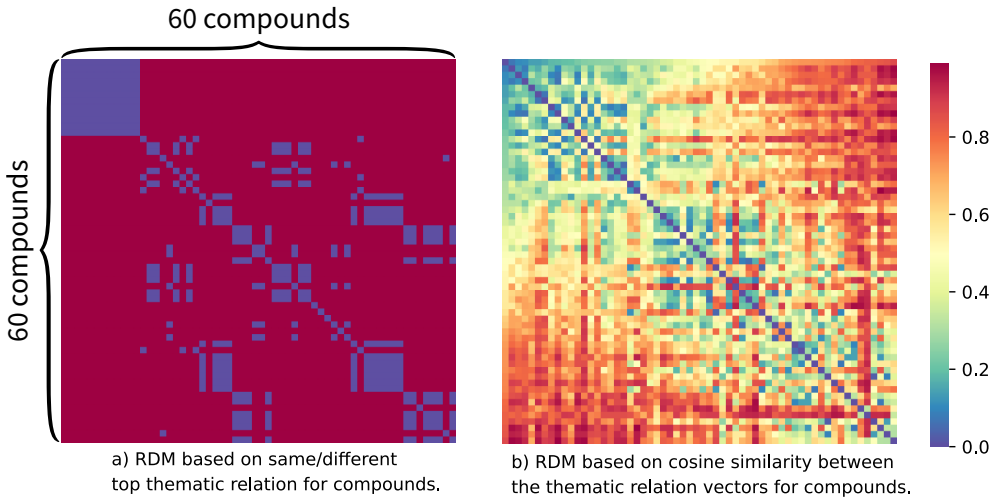

Figure 7
RDMs based on the "fine-grained" semantic relation vector representations, using 60 compounds.

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

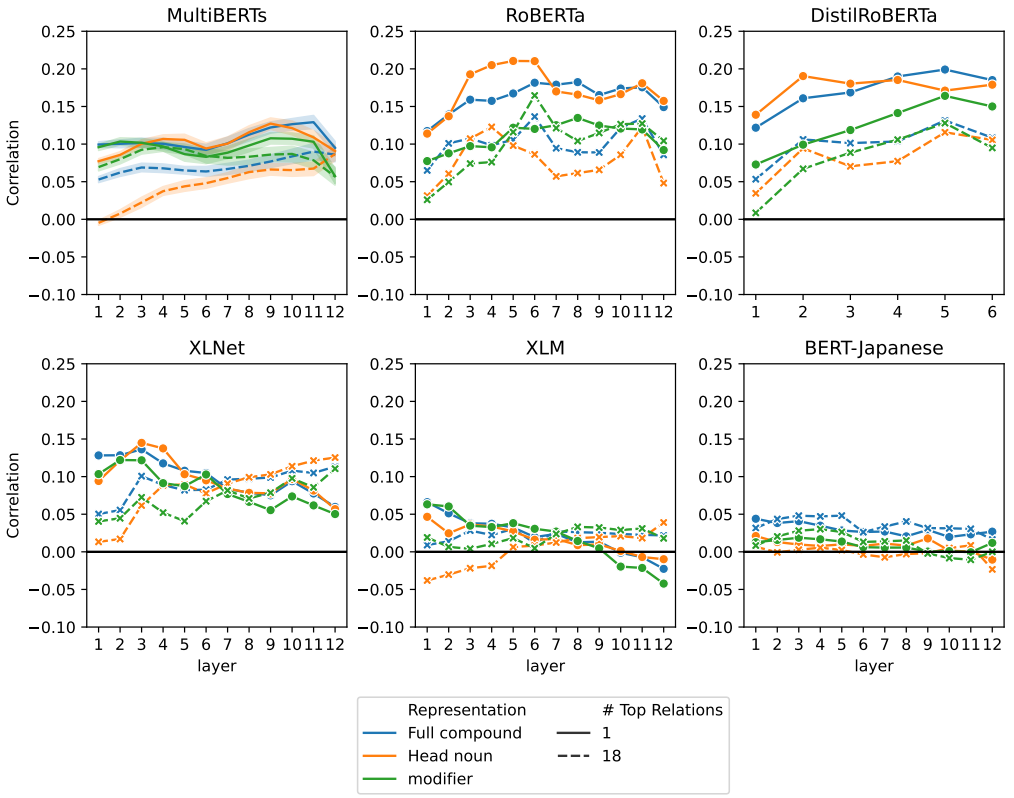

Figure 8
Results of the Relation Vector RSA experiment. Correlation between Transformer representation RDMs and the ground-truth semantic relation RDMs when (i) considering only the top mentioned thematic relation in each vector and (ii) similarity of the full 18-dimensional relation vectors.

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.