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

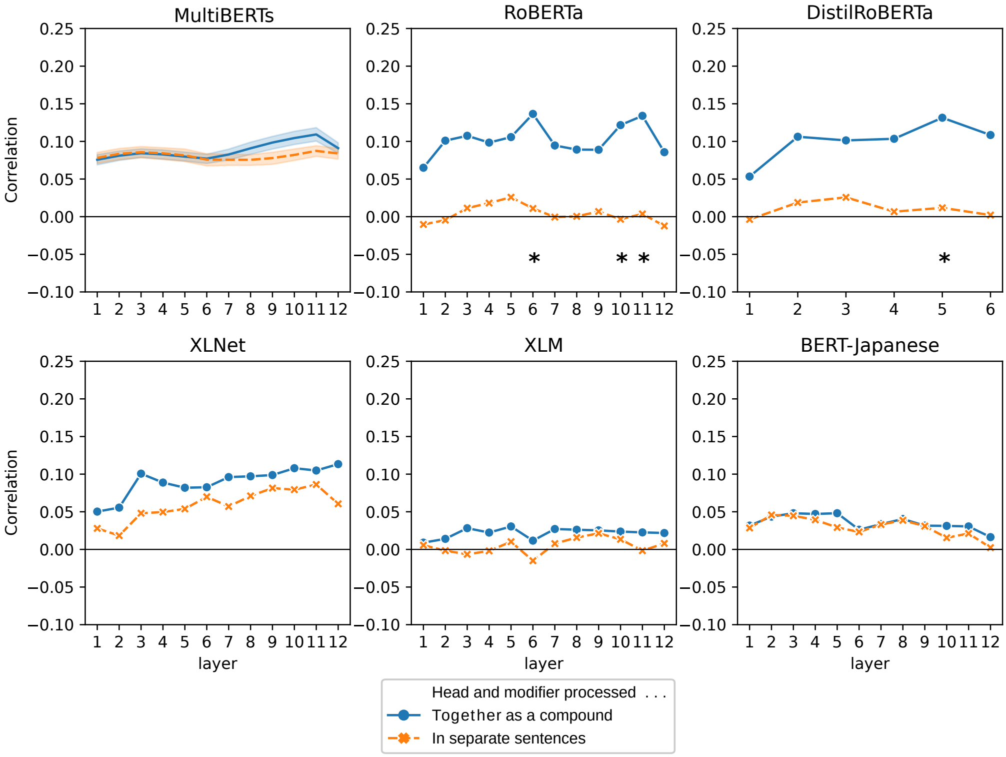

Figure 9
Results of Relation Vector and Processing Condition experiment. Correlation between Transformer representation RDMs and the ground-truth semantic relation RDM (using all 18 relations) under two processing conditions.

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