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

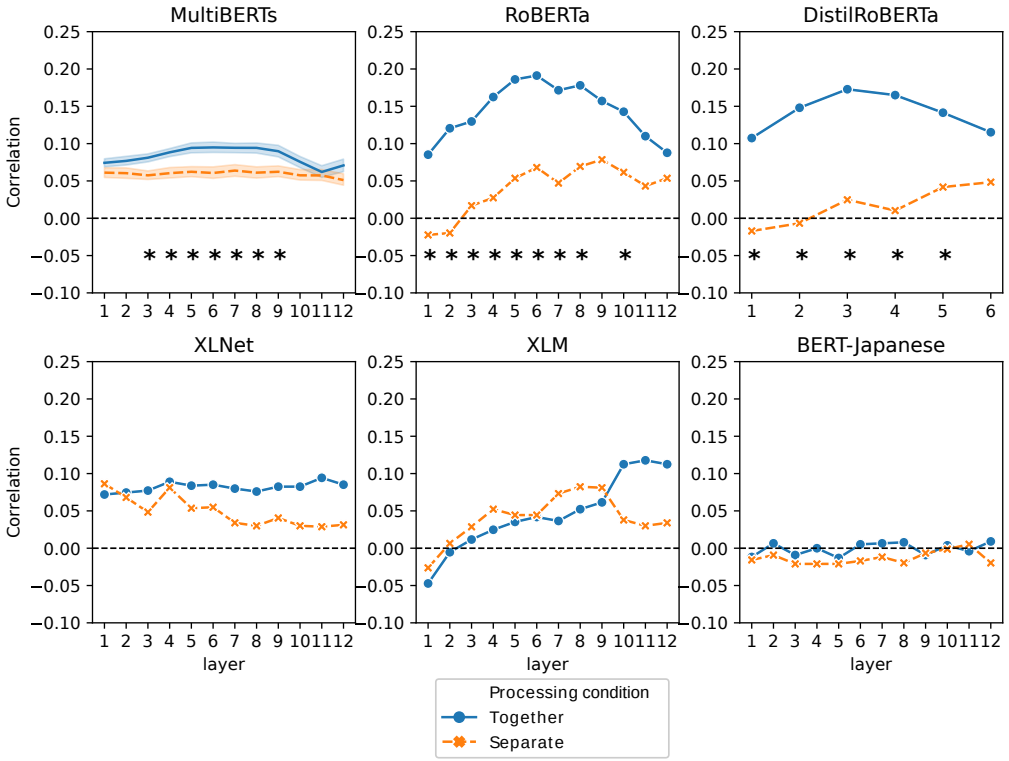

Figure 6
Results of the Relation Category and Processing Condition RSA experiment (Section 3.3). Average correlation between the same thematic relation ground-truth RDM and experimental RDMs constructed using mean-pooled token-span representations for two compound processing conditions: when the head 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