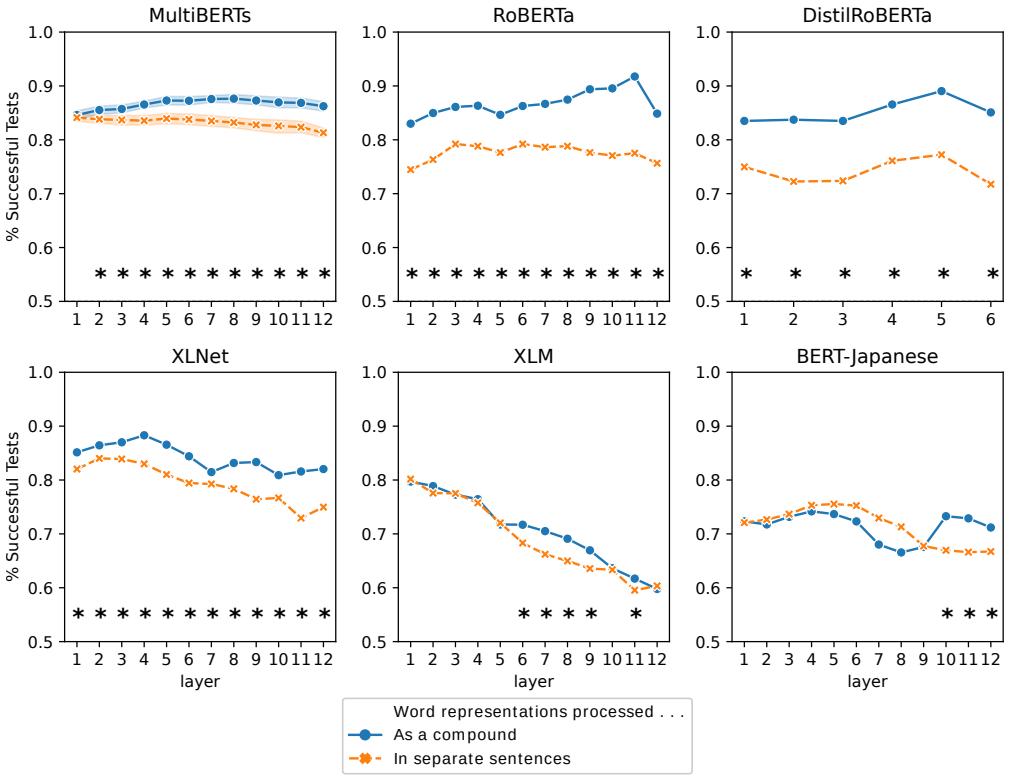

Figure 10
Results of the Compositional Probe experiment. The proportion of successes across 1770 2-vs-2 tests to decode thematic relation vectors from compound representations. Results are reported for two compound processing conditions: (i) when the head noun and modifier noun are processed together as a single compound in the same sentence, and (ii) when the head noun and modifier noun are processed in two separate sentences. Asterisks mark significant differences (at the threshold of $p < 0.05$) between the number of successes across the two processing conditions. Chance performance on this experiment is 0.5.

outcome for that pair in the Together condition is interchangeable with the observed outcome for that pair in the Separate condition. Thus, in one run of our randomization procedure, the two outcomes for each pair are randomly assigned to the two conditions, and we calculate the difference between the number of successes in the two conditions. We perform 10,000 runs of this randomization procedure, to build a distribution of these differences assuming the null hypothesis is true. Finally, we obtain a p-value for a one-sided test testing whether the actual observed number of successes for the Together condition is greater than the Separate condition, by counting the proportion of times the observed difference is greater than the differences obtained across our 10,000 randomization runs. As in the earlier experiments, 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 in Figure 10.

For all layers of the four monolingual English Transformers, the thematic relation vector is more decodable from the compound representation when the head and modifier words are processed together in the same sentence. This compositional gain from

same-context processing was found to be statistically significant in all layers of these four models excluding the first layer of the MultiBERT models. Within the four most decodable models, representations from the MultiBERT models gain the least from the normal contextual processing condition. Despite the relatively low difference in successful tests for the MultiBERT models across the two conditions, the fact that these differences are significant across the 25 instantiations of the same model suggests that this effect is robust and that BERT models consistently produce better representations of the semantic relation vector when the head and modifier words are processed together as a compound. Alongside the RSA analyses of previous experiments, these results again indicate that the Transformer-based language models that most strongly encode compound semantic relations also tend to compositionally integrate their knowledge of the head noun and modifier word in order to represent semantic relation information, above and beyond what can be decoded by relying on any association between thematic relations and the individual words.

Although the results above demonstrate better than chance decoding accuracy in the 2 vs. 2 experiment for most models and layers, we next investigated which individual compounds had poor quality predicted relation vectors, to better understand how the Transformer-based language models may fail to capture the compositional semantics of compounds. For the 2 vs. 2 decoding experiment, we save the predicted vector $\tilde{Y}^i$ of every compound i, averaging the predicted vectors across the 59 tests where compound i appears in the pair of compounds. In this way, we obtain an average relation vector prediction for each of the 60 compounds for each layer of each language model. In an exploratory analysis to investigate which compounds had generally poor quality predicted relation vectors across models and layers, we used the DBSCAN algorithm (Ester et al. 1996) to perform 60 cluster analyses—for each compound, we cluster all the predicted relation vectors, for every type of model and layer. For this analysis we select one candidate *bert-base-uncased* model from the set of MultiBERTs (as opposed to averaging the models' predictions, which would be akin to constructing an ensemble model that would perform better than any particular BERT model). We next calculated the average predicted relation vector for each cluster to obtain cluster centroids, and ranked the compounds by the greatest amount of error incurred by the best-performing cluster (i.e., the cluster with the smallest Euclidean distance between the cluster centroid and the ground-truth relation vector).

The five compounds that were most difficult to decode in the compound decoding experiment from the Compositional Probe experiment are presented in Table 2. For the compound CONSTRUCTION EQUIPMENT the ground-truth relation vector has high values for the *H for M*, *H used by M*, and *H causes M* dimensions (i.e., EQUIPMENT *for* CONSTRUCTION, EQUIPMENT *used by* CONSTRUCTION and EQUIPMENT *causes* CONSTRUCTION). However, in the clustered prediction relation vectors, the closest cluster centroid has high values for the *H derived from M*, *H made of M*, and *H is M* dimensions. Comparing the cluster's prediction for the compound CONSTRUCTION EQUIPMENT to the ground truth relation vector of STEEL EQUIPMENT, we see that the models are likely to have been overfitted for the word EQUIPMENT (i.e., equipment tends to be made of something else, but this is not actually reflected in the semantic relationship with *construction*). Similarly, the predictions for STEEL EQUIPMENT seem to be informed by the ground truth relation vector of CONSTRUCTION EQUIPMENT. We also found a similar effect for VAPOR CLOUD and VAPOR DROPS. The other two difficult compounds, CREAM CHURN and BREAKFAST SUGAR, feature unique head and modifier nouns within the dataset, and as such have not been subject to this lexical bias. For the compound CREAM CHURN, the closest cluster contains predicted relation vectors from layers 3-8 of the

Table 2
Top five most difficult compounds (as measured by the distance between predicted and actual relation vectors) on the 60 compound linear regression relation vector decoding experiment (Section 10). Compounds are ranked by greatest average distance between the best-performing set of models (grouped by clustering their predictions). The top three relation dimensions and their values are reported for the ground truth and predicted relation vectors. The models in the best-performing cluster are abbreviated. Model layer ranges are given in superscript.

Compound	Ground truth	Cluster prediction	Models in cluster
construction equipment	*H FOR M (15)* *H USED BY M* (11) *H CAUSES M (7)*	*H DERIVED FROM M* (11.4) *H MADE OF M* (9.9) *H IS M* (8.2)	BBJ^{1-12} BBU^{1-12} DB^{1-6} RB^{1-12} XBC^{1-12} $XMX1^{1-12}$
steel equipment	*H MADE OF M* (15) *H DERIVED FROM M* (14) *H IS M* (11)	*H USES M* (8.9) *H FOR M* (5.1) *H USED BY M* (5.1)	BBJ^{1-11}
vapor cloud	*H MADE OF M* (17) *H IS M* (11) *H DERIVED FROM M* (8)	*H CAUSES M* (7.8) *H MAKES M* (7.5) *H USES M* (5.1)	BBJ^{1-12} BBU^{1-12} DB^{1-6} RB^{1-12} XBC^{1-12} $XMX1^{1-8}$
cream churn	*H MAKES M* (16) *H FOR M* (10) *H USES M* (5)	*H USES M* (10.7) *M CAUSES H* (6.3) *H HAS M* (4.5)	BBJ^{3-8}
breakfast sugar	*H FOR M* (12) *H DURING M* (10) *H USED BY M* (6)	*H DERIVED FROM M* (9.2) *M MAKES H* (8.3) *M CAUSES H* (8.1)	BBU^{1-12} DB^{1-6} RB^{1-12} XBC^{1-12} $XMX1^{1-10}$

bert-base-japanese model, indicating that the English language models are producing poor representations of the relation in this compound. While CHURN *makes* CREAM is the top dimension in the ground truth, CHURN *uses* CREAM is the top predicted dimension; this may reflect the relatively few total mentions for this *H makes M* relation type across the 60 compounds (as can be seen in Figure 2). The top predictions for BREAKFAST SUGAR all erroneously indicate that sugar is derived from breakfast. One possible explanation for such predictions is that relations such as *H derived from M*, *M makes H*, and *M causes H* are commonly associated with food concepts in the dataset (e.g., OLIVE PASTE, VEGETABLE APPETIZER & GRAIN CONTROVERSY). The pattern of errors underlines the importance of controlling for associations between individual words and semantic relations when probing the models for compositional semantics, as we do in our compositional probing technique and in our contrast of the Together and Separate processing conditions in the RSA analysis.

3.6.3 Summary. All four of the monolingual models produce representations of compound nouns that are more easily decoded for head-modifier relational information when the head and modifier words are processed in the same context. This effect is stronger than we anticipated given previous work on compositionality of multi-word expressions in Transformer models. This relational information is less available for probing in the baseline Japanese and the multilingual model, which often produce