representations that are just as decodable (and sometimes less decodable) when head and modifier words are processed separately rather than as a compound.

4. Principal Findings and Implications

Using a novel approach based on ground-truth human annotations of relation meaning in the interpretation of noun-noun compounds, we used complementary representational similarity and linear probing methods to investigate whether Transformer-based language models represent the semantics of the thematic relation in noun-noun compounds. Across the experiments and analyses, we find that the English-language Transformer models, and in particular *roberta-base* and *distilroberta-base*, consistently and significantly represent compound relation semantics. Importantly, this finding persists even with careful control of the lexical content of the compounds, achieved both through psycholinguistic design that orthogonalizes relational content and the modifier and head words (i.e., the Relation Category RSA experiments) and through comparing the quality of the compound relation representation to the non-compositional representations obtained when processing the modifier and head words separately (the Relation Category/Vector and Processing Condition experiments and the Compositional Probe).

4.1 Knowledge of Implicit Intra-compound Semantic Relations

Using the 300 compound dataset and representational similarity analysis, the Relation Category RSA experiments showed that models apart from *bert-base-japanese* produce representations that moderately correlate with simple coarse-grained thematic relation signals. In the Relation Vector RSA experiments, an alternate coarse-grained RDM was constructed by considering pairs of compounds from the 60 compound dataset to be similar only if they share their top-mentioned relation. Again, this experiment provided evidence that all five Transformer-based language models that were exposed to significant English language training data (i.e., all but *bert-base-japanese*) produce token vector representations of compounds that can be distinguished by the dissimilarity pattern induced by the top-mentioned ground-truth relation (although the multilingual model, *xlm-mlm-xnli15-1024*, achieved relatively low correlations with the thematic relation signal on this task compared to the four monolingual English models).

We also found evidence that Transformer-based language models learn multi-dimensional fine-grained aspects of the semantic relation between head nouns and modifier nouns in English noun-noun compounds using the 60 compounds relation vector dataset (Devereux and Costello 2005) coupled with RSA (in the Relation Vector RSA experiments) and linear regression probing models (in the Compositional Probe experiment). In the Relation Vector RSA experiments it was found that the 18-dimensional relation vector representation task was more difficult than distinguishing compounds by whether they share their primary relation. Despite this increased difficulty, it was shown that token vector representations from four of the monolingual English models achieve moderate correlations with the 18-dimensional representational dissimilarity matrix. While *xlm-mlm-xnli15-1024* only achieved correlations of around one-third the strength of the most highly correlated model (*roberta-base*), this model demonstrated evidence of increasingly stronger correlations towards later layers. On the other hand, the baseline non-English monolingual model, *bert-base-japanese*, achieved a consistently low correlation strength, which may have been inflated slightly by a lexical overlap bias. In the Compositional Probe experiment, we found clear evidence that every model apart from the Japanese model produces token vectors that can be used to predict the 18-dimensional compound with linear regression probing models. Again, in this task the BERT-style models excelled over *xlnet-base-cased* and *xlm-mlm-xnli15-1024*. Taken together, these results clearly show that Transformer-based language models learn to encode information about the semantic relation between head nouns and modifier words in English noun-noun compounds. This finding conflicts (to an extent) with Yu and Ettinger (2020), where little evidence of compositionality was found in Transformer model representations using similarity ratings and paraphrase classifications. In contrast to that work, the present analysis uses an explicit model of compound semantics based on human annotation data for thematic relations, allowing us to directly measure the semantic representation for a given compound. However, in cases where Yu and Ettinger (2020) identify limited evidence of compositionality, they find that RoBERTa outperforms both XLM and XLNet, a result that generally aligns with our findings.

4.2 Encoding Mechanisms for Intra-compound Semantic Information

Our second major research question was how information about the semantic relation between the head and modifier word in a compound noun was encoded in Transformer-based language models. We identified three main areas of interest within this investigation: (1) whether the representation of this semantic relation results from a dynamic composition mode rather than relying on memorizing distributional co-occurrence information, (2) if this information is primarily localized within a particular token span within the compound vectors (i.e., in the head or modifier token vectors), and (3) whether this information was generally localized to a particular layer or set of layers within Transformer-based language models.

We investigated the question of whether Transformers dynamically compose information about the thematic relation between head words and modifier words by developing two types of "compositional probes" that check for statistical differences between how well a Transformer represents this semantic information under two processing conditions: (1) when a head and modifier words are processed normally as a single compound in the same context, and (2) when the head and modifier words are processed in separate sentences before their token vectors are mean-pooled. We argue that if a Transformer-based language model encodes more relational information about the compound under the same-context processing condition, then this model does not rely solely on distributional information about the co-occurrence of particular head/modifier words and their likelihood to be used with particular semantic relations; instead, they must be representing compositional relational information that is true of the compound as a phrase. In the RSA version of this compositional probe (the Relation Category and Processing Condition RSA experiment), we found that almost all layers of the English monolingual models benefited significantly from same-context processing condition. In contrast, only the final few layers of the multilingual model showed a compositional gain (a difference which was not statistically significant) and the baseline Japanese model achieved around chance levels of decodability under both conditions. Of the models that did demonstrate a significant compositional gain in our compositional probe (i.e., the normal compound processing condition), *roberta-base* and *distilroberta-base* both demonstrated the largest compositional gain and the overall best correlation to the ground-truth relation dissimilarity matrix, although the MultiBERTs models and *xlnet-base-cased* were relatively easily decodable despite benefiting less from processing constituent words of a compound in the same sentence. In the linear regression decoding version of this analysis (the Compositional Probe experiment), we found evidence that most layers of all models except *xlm-mlm-xnli15-1024* and *bert-base-japanese* benefit significantly from dynamic compositional processing in the 60-compound setting. These results generally show that Transformer-based language models produce representations of compounds that best encode for semantic information about how the head word relates to the modifier word when these constituent words are processed in the same context.

Another question of interest within the overall enquiry into how these models encode semantic information relating to head and modifier words is to what extent this information is localized to particular token spans within a Transformer's representation of a noun-noun compound. In order to shed light on this area, we included representations from head and modifier word token spans in our Relation Category and Relation Vector experiments. In these experiments we investigate the representation of both the coarse-grained and fine-grained semantic relation signal using our two main analysis techniques (RSA and linear decoding). In the Relation Category RSA experiment, it is difficult to point to any broad representational trends other than that the correlation strengths of the modifier noun vector and the full compound vector were similar in most of the Transformer-based language models we investigated. Nonetheless, the fact that the baseline Japanese model achieves a relatively high correlation with the relation type RDM in the Relation Category RSA experiment is somewhat conspicuous. In the Relation Vector RSA experiments, we observe a pattern across the BERT-style models whereby the head noun is preferred for representing the shallow top-mentioned relation RDM, whereas the modifier noun in many cases elicits stronger correlations than even the whole-compound vector. These trends however do not tend to hold for *xlm-mlm-xnli15-1024* and *xlnet-base-cased*, where the differences are more difficult to interpret.

The final aspect of how implicit intra-compound thematic relation information is encoded that we investigated is whether this information is localized to particular layers of processing. Across our experiments we find disparate trends in the results with respect to both correlation with ground-truth RDMs and decodability scores across layers. Among the BERT-style models, we find that various measures of the semantic relation signal are strongest in early/middle layers of the MultiBERT models and middle/late layers of *roberta-base* and *distilroberta-base*. In the Relation Category RSA experiment, we can see that the coarse-grained thematic relation signal of the MultiBERT models is at its strongest in the middle layers, before this information diminishes in the final few layers of processing. This result appears to contrast with Tenney, Das, and Pavlick (2019), who found that semantic information appears in later layers of BERT. We note however that performance on the semantic tasks used in that work (i.e., semantic role labeling [SRL] and coreference) reflect a model's ability to process high-level semantic information. It could be the case that a model's capacity to distinguish between words using semantic concepts such as thematic relation is a prerequisite to capturing the high-level semantic information required in SRL and coreference resolution. Furthermore, Tenney, Das, and Pavlick (2019) note that semantic information is dispersed widely across layers (compared to the stronger localization of information associated with syntactic processing). This phenomenon can be seen in the results of the four English-language monolingual models in the Relation Category RSA experiment, where we can recover a relatively good representation of the coarse-grained thematic relation signal from the least correlated layers. Interestingly, we find several cases where the layer-wise correlation trends of *xlnet-base-cased* align with those seen in *xlm-mlm-xnli15-1024*. In particular, both of these models show a coarse-grained/fine-grain trade-off in the Relation Vector RSA experiments, and both of these models decrease in decodability