2.1 Data

The first dataset (Gagné 2001) is made of 300 English noun-noun compounds that are organized into 60 groups of five compounds each. The dataset was originally compiled in order to investigate lexical and relational priming in psycholinguistic experiments on human understanding of noun-noun compounds. Using a taxonomy of 16 thematic relation types, each compound is annotated with the most appropriate thematic relation for describing the semantic relationship that exists between the head noun and the modifier (Gagné and Shoben 1997). Each group of five compounds is composed of a target compound followed by four compounds that feature (a) either a different head or modifier word from the target compound and (b) either the same or a different relation between the head noun and modifier word from the target compound, covering four different experimental conditions (see Table 1). Within each group, each modifier and head occurs with a thematic relation that is highly frequent with the modifier (e.g., the modifier MOUNTAIN often occurs with a *located in* relation, as in MOUNTAIN BREEZE but rarely occurs with an *about* relation, as in MOUNTAIN MAGAZINE). In this way, the occurrence of relation type with the individual modifier and head nouns is controlled in the experimental design (see Gagné [2001] for details).

The taxonomy of 16 thematic relation types utilized by Gagné (2001) is a useful, but rather coarse-grained, representation of the semantics of the relation instantiated for particular compounds. In many cases, several relation types may capture the meaning of a given compound, to varying degrees. We therefore also make use of a dataset of 60 compounds (a subset of the 300 compounds described above) where 34 participants rated the appropriateness of 18 different thematic relations for every compound (Devereux and Costello 2005).

These relation types and the number of total mentions for each of the types are presented in Figure 2. Compounds for which the semantic link between the head word and modifier word are similar (e.g., PROPANE STOVES and GAS LAMPS) tend to have similar distributions of appropriateness ratings across the thematic relations (see Devereux and Costello [2005] for details), and the thematic relation ratings can therefore be utilized as 18-dimensional vector representations of the relational meaning used in compounds. Relation vectors for three compounds are presented in Figure 3. Here we observe that PROPANE STOVES and GAS LAMPS are close together in the relation space (consistent with the "H uses M as fuel" relationship found in both compounds), whereas PROPANE STOVES and RAIN DROPS have very different relation vectors.

For all compounds in the datasets described above, we construct a corpus of simple, neutral sentences in the form of "It is a {*compound*}" for singular compounds (e.g.,

Table 1
Five compounds that make up one of the 60 compound groups in the Gagné (2001) noun-noun compound dataset, used in our Relation Category RSA experiments.

Modifier (M)	Head (H)	Experimental condition	Thematic relation
mountain	breeze	Target	*H LOCATED M*
kitchen	breeze	Same head noun, same relation	*H LOCATED M*
storm	breeze	Same head noun, different relation	*H DURING M*
mountain	cabin	Same modifier, same relation	*H LOCATED M*
mountain	magazine	Same modifier, different relation	*H ABOUT M*

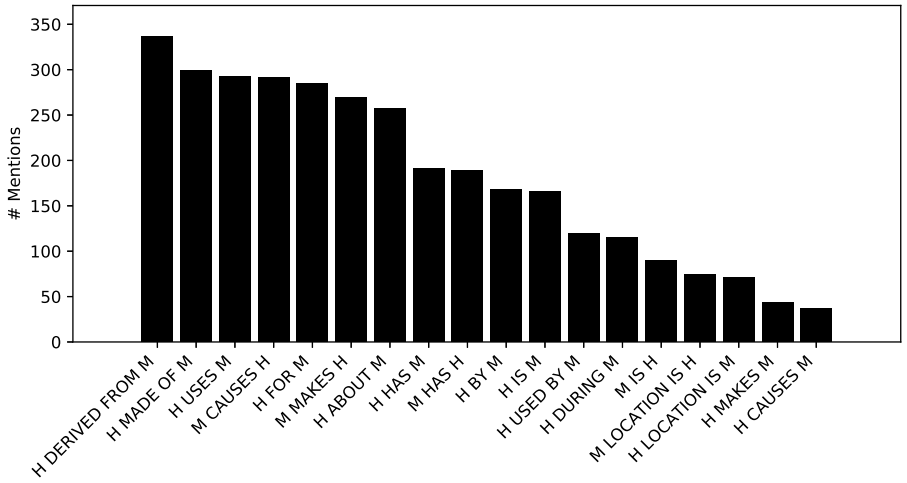

Figure 2
Distribution of relation types across the 60 compounds (Devereux and Costello 2005) dataset.

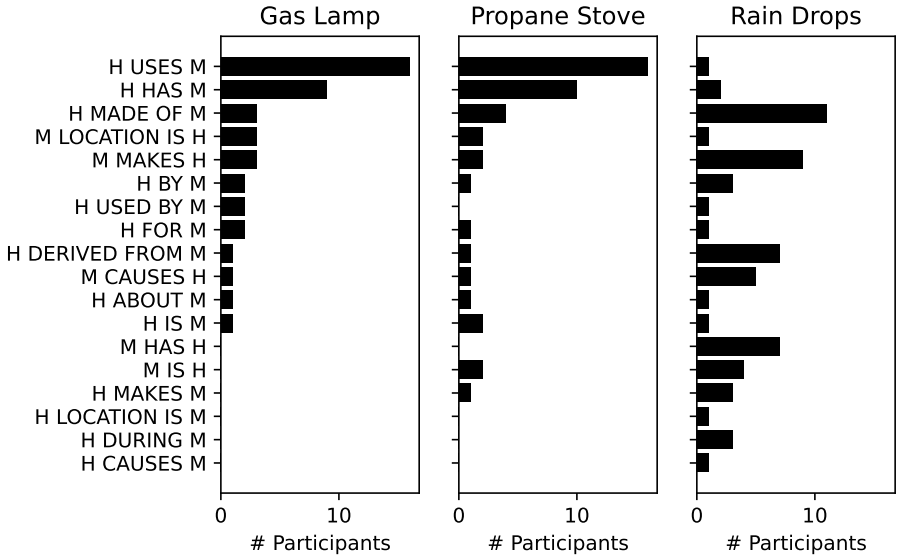

Figure 3
Sample relation vectors, based on the number of participants selecting each relation for each compound, for three compounds given in the Devereux and Costello (2005) dataset. GAS LAMP and PROPANE STOVE are close in this 18-dimensional space, reflecting the semantic similarity of the relational link (both stoves and lamps contain gas/propane that they use as fuel), while both compounds share relatively little overlap with RAIN DROPS.

"It is a wood stove"), "It is {*compound*}" for mass noun compounds (e.g., "It is solar power") and "They are{*compound*}" for plural compounds (e.g., "They are summer clothes"). The motivation for using compounds in such minimalistic sentences was to present the language models with naturalistic sentences as they would encounter

in training, while at the same time minimizing variability due to extrinsic context. As all compounds from the Devereux and Costello (2005) relation vector dataset are found in the 300 compound dataset, we use these sentences in both data settings. For our compositional analyses (the Relation Category and Processing RSA experiment [Section 3.3], the Relation Vector and Processing RSA experiment [Section 3.5], and the Compositional Probe experiment [Section 3.6]), we construct two similar sentences for all compounds using the head or modifier nouns in isolation; for example, the compound WAR RIOTS yields the sentences "It is a war" and "They are riots". After creating the corpus of minimalistic sentences, we wished to evaluate whether these constructions are particularly implausible (which could limit the generalizability of our findings) and check whether there are large disparities in sentence plausibility across different compound relations (which could potentially introduce a confound in our analysis). To this end, we used GPT-2 (Radford et al. 2019) (an autoregressive Transformer model) to calculate the perplexity of each sentence before measuring whether perplexity significantly correlated with relation magnitude for any of the relation vector dimensions. We measured an average perplexity of 267.12, compared to an average perplexity of 774.51 across sentences of similar lengths (i.e., between 5 and 14 words long) in the WikiText-2 dataset (Merity et al. 2016).[1] We then measured the Pearson's correlation between the perplexity of each sentence and the magnitude of the relation for each of the 18 dimensions in the relation vector, finding no significant correlation between any relation type and the likelihood of sentences in our corpus.

2.2 Models

In this work we considered six different Transformer-based language models (Vaswani et al. 2017). We follow the BlackboxNLP 2020 Shared Interpretation Mission (Alishahi et al. 2020) in the choice of models: BERT (bert-base-cased)[2] (Devlin et al. 2018), BERT-Japanese (bert-base-japanese)[3], RoBERTa (roberta-base) (Liu et al. 2019), Distil-RoBERTa (disilroberta-base) (Sanh et al. 2020), XLM (xlm-mlm-xnli15-1024) (Lample and Conneau 2019), and XLNet (xlnet-base-cased) (Yang et al. 2019). All of the Transformer models we target are masked language models (although *xlnet-base-cased* has been exposed to an autoregressive pre-training regime). For our BERT analysis we have made use of the MultiBERTs resource (Sellam et al. 2021), which enables us to carry out experiments on 25 different versions of the *bert-base-uncased* model that have been trained with different starting weight initializations and different shuffling of the training data, allowing for a more robust analysis of the representational trends in BERT-style models. These models were chosen to assess whether our analyses generalize over a diverse range of Transformer-based language model design choices, including monolingual/multilingual data, model size, and choice of training objective and training data. Furthermore, these models have also been used by the most relevant studies to the current work (Yu and Ettinger 2020, 2021) enabling a more direct comparison between our approach to probing Transformer models for compositional semantics and theirs.

1 Perplexity was calculated for each sentence separately before taking the average. When perplexity is calculated using the common sliding-window strategy (using preceding sentences to predict tokens), the average perplexity is 25.17 across the WikiText-2 dataset. In the sliding-window setting the model is able to leverage a large amount of contextual information to predict the recurring template structure, which would produce an extremely low perplexity score for our generated sentences.

2 Model names in the HuggingFace library are given in parentheses.

3 `https://github.com/cl-tohoku/bert-japanese`.