A Treebank dataset tree types

Due to space constraints, we only show the top 20 tree types. This can be found in Table 7.

B Treebank dataset phrase lengths

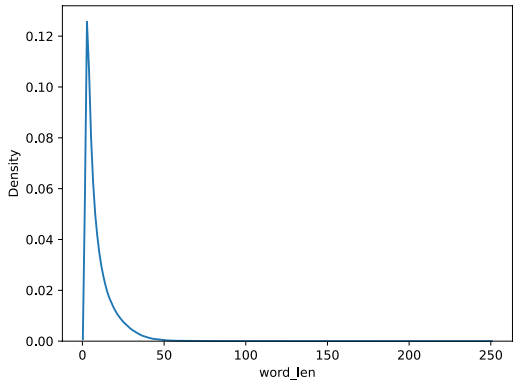

Figure 6: Length distribution of phrases mined from the treebank, in number of words. The modal length was 3 words, followed closely by 2 words. Few phrases contained more than 50 words.

C Probe learning curves

Learning curves of the approximative probes (across 10 folds) are shown in Figure 7.

D Length Correlation

The correlations of the phrase length (in words) and compositionality scores in Treebank are shown in Table 8.

E Error ratio of probes

Model/representation	Probe	Mean err. ratio (↓)
BERT_{CLS}	ADD	0.4668
	W1	0.7806
	W2	0.3903
	LIN	0.3940
	AFF	0.3908
	MLP	**0.3830**
$\text{RoBERTa}_{\text{CLS}}$	ADD	0.4152
	W1	0.7946
	W2	**0.2980**
	LIN	0.3063
	AFF	0.3013
	MLP	0.3065
$\text{DeBERTa}_{\text{CLS}}$	ADD	0.7577
	W1	**0.4661**
	W2	0.7090
	LIN	0.6777
	AFF	0.9373
	MLP	0.5856
$\text{GPT-2}_{\text{last}}$	ADD	0.4668
	W1	0.7806
	W2	**0.3903**
	LIN	0.3940
	AFF	0.3908
	MLP	0.3830
BERT_{AVG}	ADD	0.3873
	W1	0.8060
	W2	0.2167
	LIN	0.2327
	AFF	**0.2098**
	MLP	0.2283
$\text{RoBERTa}_{\text{AVG}}$	ADD	0.4504
	W1	0.8422
	W2	0.2431
	LIN	0.2471
	AFF	**0.2095**
	MLP	0.2181
$\text{DeBERTa}_{\text{AVG}}$	ADD	0.4472
	W1	0.8886
	W2	0.3202
	LIN	0.3143
	AFF	**0.3044**
	MLP	0.2952
$\text{GPT-2}_{\text{AVG}}$	ADD	0.5013
	W1	0.9074
	W2	0.4226
	LIN	0.4041
	AFF	**0.3475**
	MLP	0.3554

Table 9: Error ratio ($\frac{\text{dist}_{\text{probe}}}{\text{dist}_{\text{control}}}$) for probes trained to predict representations from different model types. Mean across 10 folds.

Tree type	Count	Example
PP → IN NP	77716	((in) (american romance))
S → NP-SBJ VP	62948	((he) (said simultaneously, "i wish they were emeralds"))
NP → DT NN	40876	((the) (way))
NP → NP PP	35743	((the temporal organization) (of the dance))
S → NP-SBJ S|<VP-.>	24467	((the partners) (said they already hold 15 % of all shares outstanding.))
VP → TO VP	21833	((to) (be the enemy))
PP-LOC → IN NP	18005	((in) (the marketplace))
NP → DT NP|<JJ-NN>	14898	((a) (professional linguist))
VP → MD VP	13575	((could) (make up his mind))
VP → VB NP	11838	((evaluate) (the progress of therapy))
PP-TMP → IN NP	11032	((for) (almost a year))
PP-CLR → IN NP	10054	((from) (the most sympathetic angle))
NP → NNP NNP	9863	((honolulu) (harbor))
NP → JJ NNS	9477	((recent) (years))
VP → VBD VP	8356	((was) (salted))
SBAR → WHNP-1 S	8332	((what) (to look for))
SBAR → IN S	7848	((that) (it exceeds the company 's annual sales and its market capitalization))
NP-SBJ → DT NN	7600	((the) (rebound))
S → NP-SBJ-1 VP	7486	(draperies) (could be designed to serve structural purposes)
NP → NP SBAR	7317	((the " culture shock ") (they might encounter in remote overseas posts))

Table 7: Counts of the top 20 grammatical tree types found in the WSJ and Brown sections of the Penn Treebank, with some examples given.

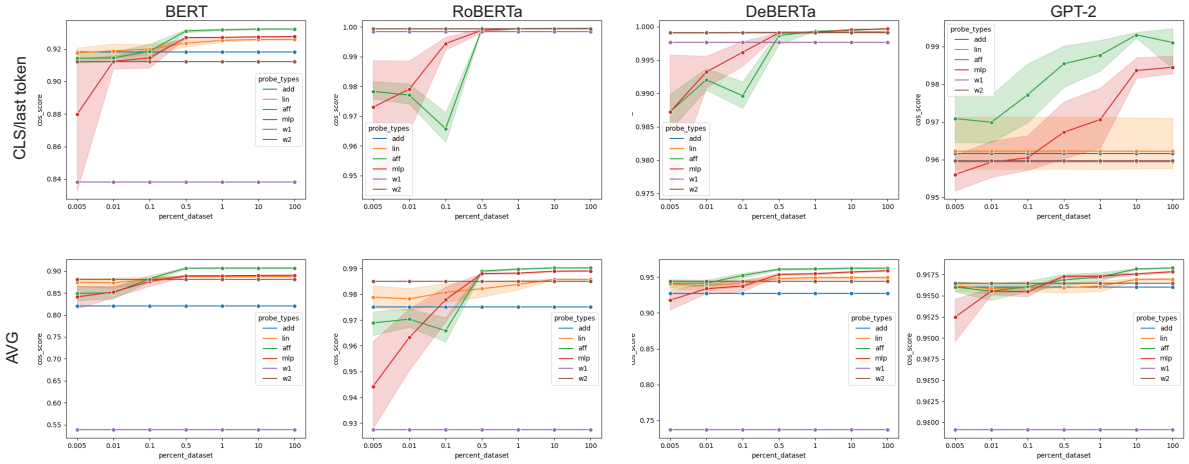

Figure 7: Learning curves of approximative probes trained on differing percentages of train data.

Model and representation	Spearman ρ	p-val
BERT_{CLS}	-0.0700	0.0
$\text{RoBERTa}_{\text{CLS}}$	0.1659	0.0
$\text{DeBERTa}_{\text{CLS}}$	0.1166	0.0
BERT_{AVG}	0.7143	0.0
$\text{RoBERTa}_{\text{AVG}}$	0.7086	0.0
$\text{DeBERTa}_{\text{AVG}}$	0.7866	0.0

Table 8: Spearman ρ correlation between phrase length (in words) and compositionality score in the treebank.

F Annotation setup and instructions

Annotators were recruited from a population of graduate students. Initially, 6 annotators completed the pilot experiment, which consisted of 101 examples. The subset of three annotators with highest agreement was asked if they would like to complete the full study. One annotator in the highest-agreement group could not continue to the full study, so this annotator was excluded, and the next group with highest agreement was chosen. The agreement values in subsubsection 6.1.1 are for the final group of annotators chosen.

The experiment was implemented on the Qualtrics platform, and participants were first presented with a consent form, linking to more background information on the study, and informing them that their participation was entirely voluntary. After agreeing to the terms, participants were shown some examples and went through 3 practice questions. The example given are shown in Figure 8, and the annotation interface is shown in Figure 9 and Figure 10. After completing the practice section, annotators began annotating the real examples, which followed the same interface as the practice examples.

Annotators were all located in the United States, paid approximately $15 per hour for their work.

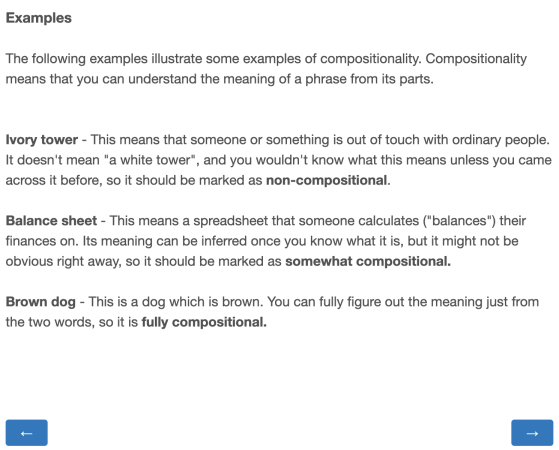

Figure 8: Examples of compositionality judgments shown to annotators

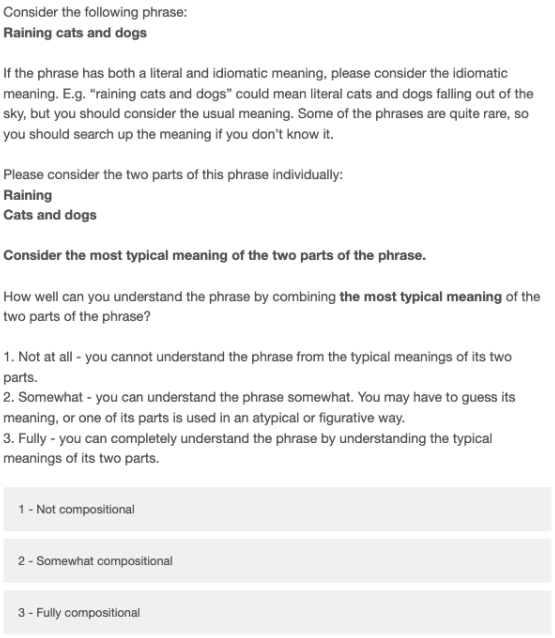

Figure 9: First page of annotation interface for a practice phrase

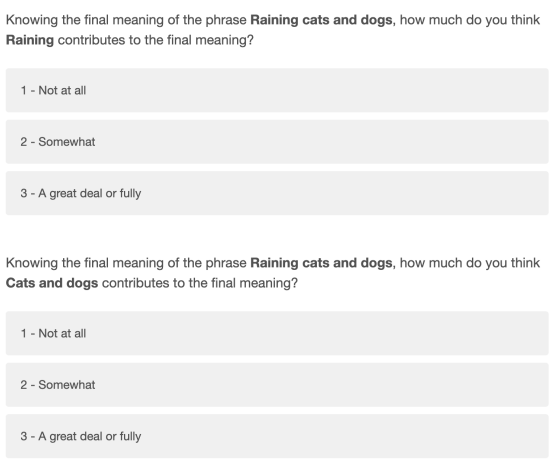

Figure 10: Second page of annotation interface for a practice phrase