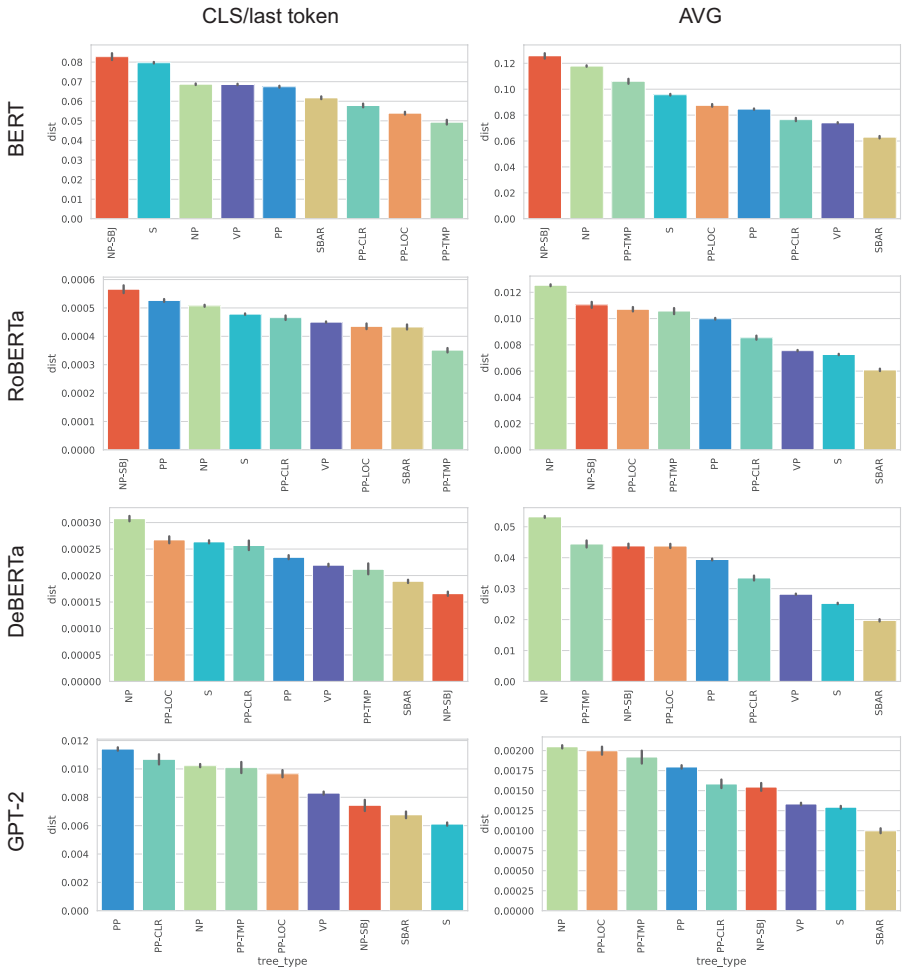

Figure 5: Tree reconstruction error (cosine distance) for each parent phrase type present in the Treebank, ordered from highest mean error to lowest. Based on the affine approximation for each model and representation type. Expanded version with all tree types is presented in Appendix I.

Phrase	Idiom	Judgment	Subphrase contribution
Making heavy weather	Yes	1 - Not compositional	Making: 1 - Not at all Heavy weather: 2 - Somewhat
Chief part	No	2 - Somewhat compositional	Chief: 2 - Somewhat Part: 3 - A great deal
Portrait of Washington	No	3 - Fully compositional	Portrait: 3 - A great deal of Washington: 3 - A great deal

Table 3: Example judgments of one annotator on the pilot set. Annotators were asked to rate each phrase from 1 to 3, where 1 meant not compositional and 3 meant fully compositional. They were also asked how much each subphrase contributed to the meaning.

6.1.2 Model Comparison

To compare human judgments to model compositionality scores, we use the best trained approximative probe for each model and representation type to predict a vector for the full phrase based on its left and right subphrases (taking the probe

trained on the first fold). We use cosine similarity to the expected representation as the measure of how compositional a phrase is for a model and representation type.

We take the Spearman correlation between model compositionality scores and human compositionality judgments and observe differences between human judgments and compositionality scores from model representations.

6.2 Results

6.2.1 Correlation with human judgments

There is a weak correlation between model and human compositionality scores. The most promising trend is found in RoBERTa, where both **CLS** and **AVG** representations have a significant positive correlation with human judgments. Results are in Table 4, with corrected p-values (Holm, 1979).

Model and representation	Spearman ρ	p-val
BERT_{CLS}	-0.02308	0.9915
$\text{RoBERTa}_{\text{CLS}}$	0.1913	9.7934×10^{-8}*
$\text{DeBERTa}_{\text{CLS}}$	0.01466	0.9915
$\text{GPT-2}_{\text{last}}$	0.009428	0.02654*
BERT_{AVG}	0.1283	8.594×10^{-4}*
$\text{RoBERTa}_{\text{AVG}}$	0.1386	2.782×10^{-4}*
$\text{DeBERTa}_{\text{AVG}}$	-0.03819	0.7792
$\text{GPT-2}_{\text{AVG}}$	-0.04598	0.6987

Table 4: Spearman correlation between human judgments of compositionality and compositionality score generated by different model and representation combinations. P-values are corrected for multiple comparisons with the Holm-Bonferroni correction.

6.2.2 Subphrase Contribution Test

Annotators indicated to what extent they believed each part of the phrase contributed to the final meaning. We examined examples in which annotators rated one part of the phrase, for example a, as contributing more to the final meaning, and checked how often $d_{cos}(r(x), r(a)) > d_{cos}(r(x), r(b))$. Models do surprisingly poorly at this test, with most performing below chance. Results are presented in Table 5. An error analysis on $\text{RoBERTa}_{\text{AVG}}$ indicated that in many cases, errors were due to idiomaticity failures. For example, "noble gas" is a type of gas that was rated as being more similar to "gas" by humans, but "noble" by RoBERTa.[7]

[7]Similar errors were made for phrases such as "grandfather clock", "as right as rain", "ballpark estimate". A "grandfather clock" is a type of clock, "as right as rain" indicates that something is alright, and a "ballpark estimate" is a rough estimate.

Model and representation	Subphrase accuracy
BERT_{CLS}	49.71%
$\text{RoBERTa}_{\text{CLS}}$	45.91%
$\text{DeBERTa}_{\text{CLS}}$	45.61%
$\text{GPT-2}_{\text{last}}$	43.86%
BERT_{AVG}	52.92%
$\text{RoBERTa}_{\text{AVG}}$	45.03%
$\text{DeBERTa}_{\text{AVG}}$	46.20%
$\text{GPT-2}_{\text{AVG}}$	45.32%
	Idiomatic accuracy
BERT_{CLS}	45.60%
$\text{RoBERTa}_{\text{CLS}}$	60.03%
$\text{DeBERTa}_{\text{CLS}}$	56.67%
$\text{GPT-2}_{\text{last}}$	59.15%
BERT_{AVG}	57.57%
$\text{RoBERTa}_{\text{AVG}}$	58.98%
$\text{DeBERTa}_{\text{AVG}}$	45.77%
$\text{GPT-2}_{\text{AVG}}$	48.42%

Table 5: Accuracy of model representations on the subphrase test and idiomaticity test.

6.2.3 Idiomaticity Test

Because idioms were matched with non-idiomatic expressions, we tested for correctly identifying the idioms. We limited the analysis to pairs where the idiomatic expression was rated as less compositional than the matched expression. Results are shown in Table 5. Results are better than the subphrase contribution test, but models do not achieve good results, the best performing representation being $\text{RoBERTa}_{\text{CLS}}$.

6.2.4 Correlations with Other Factors

We examine correlations of model and human compositionality scores with the frequency and length of the phrase in words. As noted before, there is a strong correlation between length and compositionality score in models but not in human results. Results are in Appendix K. A comparison of phrases rated as most and least compositional by humans, as well as RoBERTa, is presented in Table 6.

7 Related work

7.1 Background on Compositionality

Compositionality has been debated in the philosophy of language, with opposing views (Herbelot, 2020): the *bottom-up* view that the meaning of a larger phrase is a function of the meaning of its parts (Cresswell, 1973), and the *top-down* view

Model & representation	Most compositional	Least compositional
Human	"population growth" "few weeks away" "railroad monopoly"	"gravy train" "shrinking violet" "revolving door syndrome"
RoBERTa$_{CLS}$	"two small sticks" "dark glass bottle" "annual music festival"	"worse than none" "cases apart" "arch'd eyebrow"
RoBERTa$_{AVG}$	"look with open eyes" "be of equal importance" "come after breakfast"	"advertisement revenue" "taking it upon oneself" "all paces"

Table 6: Most and least compositional phrases in CHIP by human judgments and RoBERTa compositionality scores. Human scores are the average of 3 annotators.

that smaller parts only have meaning as a function of the larger phrase (Fodor and LePore, 1992). It is likely that there is a blend of bottom-up and top-down processing corresponding to compositional and non-compositional phrases respectively (Dankers et al., 2022a).

Hupkes et al. have proposed several compositionality tests based on previous interpretations: (Hupkes et al., 2020). We focus on localism, corresponding to the bottom-up view.

7.2 Other Definitions of Compositionality

Other works do other tests for compositionality, notably substitutivity (Hupkes et al., 2020). Evidence suggests that models may be unable to modulate the bottom-up and top-down processing of phrases (Dankers et al., 2022b,a). Substitutivity effects appear to not be represented well (Garcia et al., 2021; Yu and Ettinger, 2020). This indicates that phrases are not being composed as expected and motivates our study of how local composition is carried out in these models, and which types of phrase are processed top-down and bottom-up.

7.3 Studies of Localism

Previous studies of local composition focus on bigrams, particularly adjective-noun and noun-noun bigrams (Nandakumar et al., 2019; Cordeiro et al., 2019; Salehi et al., 2015; Reddy et al., 2011; Mitchell and Lapata, 2010). However, many of these studies assume an additive composition function or only fit a composition function on the bigrams in their datasets.

A study finds some evidence for successful local composition in the case of mathematical expressions, but used a constrained test set on a domain that is expected to be perfectly locally compositional (Russin et al., 2021).

7.4 Approximating LM Representations

There has been recent interest in understanding the compositionality of continuous representations generated by neural models (Smolensky et al., 2022). LM representations have been approximated as the output of explicitly compositional networks based on tensor products (McCoy et al., 2020, 2019; Soulos et al., 2020). These are typically evaluated based on compositional domains, such as the SCAN dataset (Lake and Baroni, 2017).

Previous work on the geometry of word embeddings within a sentence shows that language models can encode hierarchical structure (Coenen et al., 2019; Manning et al., 2020; Jawahar et al., 2019). However, it is an open question as to why LMs do not tend to generalize well compositionally (Lake and Baroni, 2017; Keysers et al., 2020).

8 Conclusion

We analyze the compositionality of representations from several language models and find that there is an effective affine approximation in terms of a phrase's syntactic children for many phrases. Although LM representations may be surprisingly predictable, we find that human compositionality judgments do not align well with how LM representations are structured.

In this work, we study the representations produced after extensive training. However, the consistency of several trends we observed suggests that there may be theoretical reasons why LM representations are structured in certain ways. Future work could investigate the evolution of compositionality through training, or motivate methods that would allow LMs to achieve improved compositional generalization while representing non-compositionality.

Acknowledgments

Thank you to Amanda Bertsch, Ting-Rui Chiang, Varun Gangal, Perez Ogayo, and Zora Wang for participating in compositionality annotations. This work was supported in part by a CMU Presidential Fellowship to the first author, and the Tang Family AI Innovation Fund.