Are Representations Built from the Ground Up? An Empirical Examination of Local Composition in Language Models

Emmy Liu and **Graham Neubig**
Language Technologies Institute
Carnegie Mellon University
{mengyan3, gneubig}@cs.cmu.edu

Abstract

Compositionality, the phenomenon where the meaning of a phrase can be derived from its constituent parts, is a hallmark of human language. At the same time, many phrases are *non-compositional*, carrying a meaning beyond that of each part in isolation. Representing both of these types of phrases is critical for language understanding, but it is an open question whether modern language models (LMs) learn to do so; in this work we examine this question. We first formulate a problem of predicting the LM-internal representations of longer phrases given those of their constituents. We find that the representation of a parent phrase can be predicted with some accuracy given an affine transformation of its children. While we would expect the predictive accuracy to correlate with human judgments of semantic compositionality, we find this is largely *not* the case, indicating that LMs may not accurately distinguish between compositional and non-compositional phrases. We perform a variety of analyses, shedding light on when different varieties of LMs do and do not generate compositional representations, and discuss implications for future modeling work.[1]

1 Introduction

Compositionality is argued to be a hallmark of linguistic generalization (Szabó, 2020). However, some phrases are non-compositional, and cannot be reconstructed from individual constituents (Dankers et al., 2022a). Intuitively, a phrase like "I own cats and dogs" is locally compositional, whereas "It's raining cats and dogs" is not. Therefore, any representation of language must be easily composable, but it must also correctly handle cases that deviate from compositional rules.

Both lack (Hupkes et al., 2020; Lake and Baroni, 2017) and excess (Dankers et al., 2022b) of compositionality have been cited as common sources of errors in NLP models, indicating that models may handle phrase composition in an unexpected way.

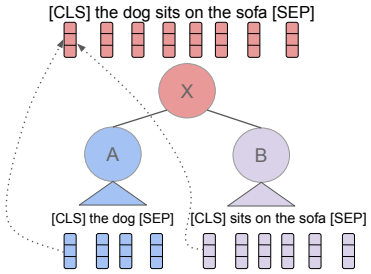

Figure 1: An illustration of the local composition prediction problem with [CLS] representations.

In general form, the compositionality principle is simply "the meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined" (Pelletier, 1994). However, this definition is underspecified (Partee, 1984). Recent efforts to evaluate the compositional abilities of neural networks have resulted in several testable definitions of compositionality (Hupkes et al., 2020).

Previous work on compositionality in natural language focuses largely on the definition of **substitutivity**, by focusing on changes to the constituents of a complex phrase and how they change its representation (Dankers et al., 2022a; Garcia et al., 2021; Yu and Ettinger, 2020). The definition we examine is **localism**: whether or not the representation of a complex phrase is derivable only from its local structure and the representations of its immediate "children" (Hupkes et al., 2020). A similar concept has been proposed separately to measure the compositionality of learned representations, which we use in this work (Andreas, 2019). We focus on localism because it is a more direct definition and does not rely on the collection of contrastive pairs of phrases. This allows us to examine a wider range of phrases of different types and lengths.

In this paper, we ask whether reasonable compo-

[1]Code and data available at https://github.com/nightingal3/lm-compositionality

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9053–9073
December 7-11, 2022

sitional probes can predict an LM's representation of a phrase from its children in a syntax tree, and if so, which kinds of phrase are more or less compositional. We also ask whether this corresponds to human judgements of compositionality.

We first establish a method to examine local compositionality on phrases through probes that try to predict the representation of a parent given its children (section 2). We create two English-language datasets upon which to experiment: a large-scale dataset of 823K phrases mined from the Penn Treebank, and a new dataset of idioms and paired non-idiomatic phrases for which we elicit human compositionality judgements, which we call the **C**ompositionality of **H**uman-annotated **I**diomatic **P**hrases dataset (**CHIP**) (section 3).

For multiple models and phrase types, we find that phrase embeddings across models and representation types have a fairly predictable affine compositional structure based on embeddings of their constituents (section 4). We find that there are significant differences in compositionality across phrase types, and analyze these trends in detail, contributing to understanding how LMs represent phrases (section 5). Interestingly, we find that human judgments do not generally align well with the compositionality level of model representations (section 6). This implies there is still work to be done at the language modelling level to capture a proper level of compositionality in representations.

2 Methods and Experimental Details

2.1 Tree Reconstruction Error

We follow Andreas (2019) in defining deviance from compositionality as *tree reconstruction error*. Consider a phrase $x = [a][b]$, where a and b can be any length > 0. Assume we always have some way of knowing how x should be divided into a and b. Assume we also have some way of producing representations for x, a, and b, which we represent as a function r. Given representations $r(x)$, $r(a)$ and $r(b)$, we wish to find the function which most closely approximates how $r(x)$ is constructed from $r(a)$ and $r(b)$.

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \delta_{x,ab} \tag{1}$$

$$\delta_{x,ab} = d(r(x), f(r(a), r(b)) \tag{2}$$

Where $\mathcal{X}$ is the set of possible phrases in the language that can be decomposed into two parts, $\mathcal{F}$ is the set of functions under consideration, and d is a distance function. An example scenario is depicted in Figure 1.

For d, we use cosine distance as this is the most common function used to compare semantic vectors. The division of x into a and b is specified by syntactic structure (Chomsky, 1959). Namely, we use a phrase's annotated constituency structure and convert its constituency tree to a binary tree with the right-factored Chomsky Normal Form conversion included in NLTK (Bird and Loper, 2004).

2.2 Language Models

We study representations produced by a variety of widely used language models, specifically the `base-(uncased)` variants of Transformer-based models: **BERT**, **RoBERTa**, **DeBERTa**, and **GPT-2** (He et al., 2021; Liu et al., 2019; Devlin et al., 2019; Radford et al., 2019).

2.2.1 Representation extraction

Let $[x_0, ..., x_N]$ be a sequence of $N + 1$ input tokens, where x_0 is the `[CLS]` token if applicable, and x_N is the end token if applicable. Let $[h_0^{(i)}, ..., h_N^{(i)}]$ be the embeddings of the input tokens after the i-th layer.

For models with the `[CLS]` beginning of sequence token (BERT, RoBERTa, and DeBERTa), we extracted the embedding of the `[CLS]` token from the last layer, which we refer to as the **CLS** representation. For GPT-2, we extracted the last token, which serves a similar purpose. This corresponds to $h_0^{(12)}$ and $h_N^{(12)}$ respectively.

Alternately, we also averaged all embeddings from the last layer, including special tokens. We refer to this as the **AVG** representation.

$$\frac{1}{N+1} \sum_{i=0}^{N+1} h_i^{(12)} \tag{3}$$

2.3 Approximating a Composition Function

To use this definition, we need a composition function $\hat{f}$. We examine choices detailed in this section.

For parameterized probes, we follow the probing literature in training several probes to predict a property of the phrase given a representation of the phrase. However, in this case, we are not predicting a categorical attribute such as part of speech. Instead, the probes that we use aim to predict the

parent representation $r(x)$ based on the child representations $r(a)$ and $r(b)$. We call this an *approximative probe* to distinguish it from the usual use of the word probe.

2.3.1 Arithmetic Probes

In the simplest probes, the phrase representation $r(x)$ is computed by a single arithmetic operation on $r(a)$ and $r(b)$. We consider three arithmetic probes:[2]

$$\textbf{ADD}(r(a), r(b)) = r(a) + r(b) \quad (4)$$

$$\textbf{W1}(r(a), r(b)) = r(a) \quad (5)$$

$$\textbf{W2}(r(a), r(b) = r(b) \quad (6)$$

2.3.2 Learned Probes

We consider three types of learned probes. The linear probe expresses $r(x)$ as a linear combination of $r(a)$ and $r(b)$. The affine probe adds a bias term. The MLP probe is a simple feedforward neural network with 3 layers, using the ReLU activation.

$$\textbf{LIN}(r(a), r(b)) = \alpha_1 r(a) + \alpha_2 r(b) \quad (7)$$

$$\textbf{AFF}(r(a), r(b)) = \alpha_1 r(a) + \alpha_2 r(b) + \beta \quad (8)$$

$$\textbf{MLP}(r(a), r(b)) = W_3 h_2 \quad (9)$$

Where

$$h_1 = \sigma(W_1[r(a); r(b)])$$

$$h_2 = \sigma(W_2 h_1),$$

W_1 is (300×2), W_2 is (768×300), and W_3 is (1×768). We do not claim that this is the best MLP possible, but use it as a simple architecture to contrast with the linear models.

3 Data and Compositionality Judgments

3.1 Treebank

To collect a large set of phrases with syntactic structure annotations, we collected all unique subphrases (≥ 2 words) from WSJ and Brown sections of the Penn Treebank (v3) (Marcus et al., 1993). [3]

The final dataset consists of **823K** phrases after excluding null values and duplicates. We collected the length of the left child in words, the length of the right child in words, and the tree's production rule, which we refer to as *tree type*. There were 50260 tree types in total, but many of these are unique. Examples and phrase length distribution can be found in Appendix A, and Appendix B.

3.2 English Idioms and Matched Phrase Set

Previous datasets center around notable bigrams, some of which are compositional and some of which are non-compositional (Ramisch et al., 2016b; Reddy et al., 2011). However, there is a positive correlation between bigram frequency and human compositionality scores in these datasets, which means that it is unclear whether models are capturing compositionality or merely frequency effects if they correlate well with the human scores.

Because models are likely more sensitive to surface features of language than humans, we gathered a more controlled set of phrases to compare with human judgments.

Since non-compositional phrases are somewhat rare, we began with a set of seed idioms and bigrams from previous studies (Jhamtani et al., 2021; Ramisch et al., 2016b; Reddy et al., 2011). We used idioms because they are a common source of non-compositional phrases. Duplicates after lemmatization were removed.

For each idiom, we used Google Syntactic NGrams to find three phrases with an identical part of speech and dependency structure to that idiom, and frequency that was as close as possible relative to others in Syntactic Ngrams (Goldberg and Orwant, 2013).[4] For example, the idiom "sail under false colors" was matched with "distribute among poor parishioners". More examples can be found in Table 1. An author of this paper inspected the idioms and removed those that were syntactically analyzed incorrectly or offensive.

4 Approximating a Composition Function

4.1 Methods

To approximate the composition functions of models, we extract the **CLS** and **AVG** representations from each model on the Treebank dataset. We used 10-fold cross-validation and trained the learned probes on the 90% training set in each fold. The

[2] Initially, we considered the elementwise product $\textbf{PROD}(r(a), r(b)) = r(a) \odot r(b)$, but found that it was an extremely poor approximation.

[3] We converted the trees to Chomsky Normal Form with right-branching using NLTK (Bird and Loper, 2004). We note that not all subtrees are syntactically meaningful. However, we used this conversion to standardize the number of children and formatting. We exclude phrases with a null value for the left or right branch (Bies et al., 1995).

[4] The part of speech/dependency pattern for each idiom was taken to be the most common pattern for that phrase in the dataset

Idiom	Matched phrase	Syntactic pattern	Log frequency
Devil's advocate	Baker's town	JJ/dep/2 NN/pobj/0	2.398
Act of darkness	Abandonment of institution	NN/dobj/0 IN/prep/1 NN/pobj/2	4.304
School of hard knocks	Field of social studies	NN/pobj/0 IN/prep/1 JJ/amod/4 NNS/pobj/2	6.690

Table 1: Examples of idioms with their matched phrases, selected based on having the same syntactic pattern and most similar log frequency in the Syntactic Ngrams dataset. Examples depicted here have the same log frequency. Note that the frequency is based on the most common dependency and constituency pattern found in Syntactic NGrams. Humans were asked to rate each phrase for its compositionality.

remaining 10% were divided into a test set (5%) and dev set (5%).[5]

To fairly compare probes, we used minimum description length probing (Voita and Titov, 2020).This approximates the length of the online code needed to transmit both the model and data, which is related to the area under the learning curve. Specifically, we recorded average cosine similarity of the predicted vector and actual vector on the test set while varying the size of the training set from 0.005% to 100% of the original.[6] We compare the AUC of each probe under these conditions to select the most parsimonious approximation for each model.

4.2 Results

We find that **affine probes** are best able to capture the composition of phrase embeddings from their left and right subphrases. A depiction of probe performance at approximating representations across models and representation types is in Figure 2. However, we note that scores for most models are very high, due to the anisotropy phenomenon. This describes the tendency for most embeddings from pretrained language models to be clustered in a narrow cone, rather than distributed evenly in all directions (Li et al., 2020; Ethayarajh, 2019). We note that it is true for both word and phrase embeddings.

Since we are comparing the probes to each other relative to the same anisotropic vectors, this is not necessarily a problem. However, in order to compare each probe's performance compared to chance, we correct for anisotropy using a control task. This task is using the trained probe to predict a random phrase embedding from the set of treebank phrase embeddings for that model, and recording the distance between the compositional probe's prediction and the random embedding. This allows us to calculate an error ratio $\frac{\text{dist}_{\text{probe}}}{\text{dist}_{\text{control}}}$, where dist_{probe} represents the original average distance from the true representation, and $\text{dist}_{\text{control}}$ is the average distance on the control task. This quantifies how much the probe improves over a random baseline that takes anisotropy into account, where a smaller value is better. These results can be found in Appendix E. The results without anisotropy correction can be found in Appendix G. In most cases, the affine probe still performs the best, so we continue to use it for consistency on all the model and representation types.

We also compare the AUC of training curves for each probe and find that the affine probe remains the best in most cases, except $\text{RoBERTa}_{\text{CLS}}$ and $\text{DeBERTa}_{\text{CLS}}$. Training curves are depicted in Appendix C. AUC values are listed in Appendix H.

Interestingly, there was a trend of the right child being weighted more heavily than the left child, and each model/representation type combination had its own characteristic ratio of the left child to the right child. For instance, in BERT, the weight on the left child was 12, whereas it was 20 for the right child.

For example, the approximation for the phrase "green eggs and ham" with BERT `[CLS]` embeddings would be: $r_{CLS}(\text{"green eggs and ham"}) = 12r_{CLS}(\text{"green eggs"}) + 20r_{CLS}(\text{"and ham"}) + \beta$.

[5]The learned probes were trained with early stopping on the dev set with a patience of 2 epochs, up to a maximum of 20 epochs. The Adam optimizer was used, with a batch size of 512 and learning rate of 0.512.

[6]We look at milestones of 0.005%, 0.01%, 0.1%, 0.5%, 1%, 10% and 100% specifically. This was because initial experimentation showed that probes tended to converge at or before 10% of the training data. Models were trained separately (with the same seed and initialization) for each percentage of the training data, and trained until convergence for each data percentage condition.

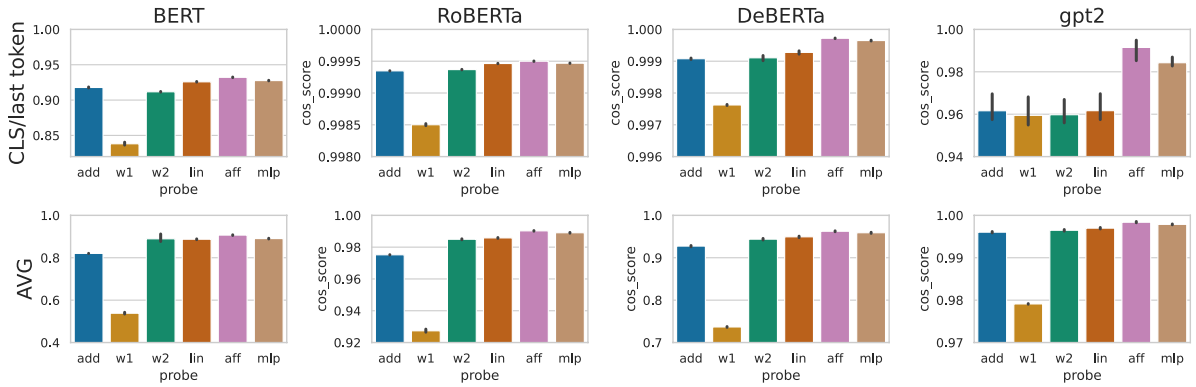

Figure 2: Mean compositionality score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Error bar indicates 95% CI.

5 Examining Compositionality across Phrase Types

5.1 Methods

Intuitively, we expect the phrases whose representations are close to their predicted representation to be more compositional. We call similarity to the expected representation, $\text{sim}(r(x), \hat{f}(r(a), r(b)))$, the *compositionality score* of a phrase.

We record the mean reconstruction error for each tree type and report the results. In addition to comparing tree types to each other, we also examine the treatment of named entities in subsubsection 5.2.1. We examine the relationship between length of a phrase in words and its compositionality score in subsubsection 5.2.2.

5.2 Results

There is a significant difference between the mean compositionality score of phrase types. Particularly, the **AVG** representation assigns a lower compositionality score to NP $\rightarrow$ NNP NNP phrases, which is expected since this phrase type often corresponds to named entities. By contrast, the **CLS** representation assigns a low compositionality score to NP $\rightarrow$ DT NN, which is unexpected given that such phrases are generally seen as compositional. The reconstruction error for the most common phrase types is shown in Figure 5.

Because different phrase types may be treated differently by the model, we examine the relative compositionality of phrases within each phrase type. Examples of the most and least compositional phrases from several phrase types are shown in Table 2 for RoBERTa$_{\text{CLS}}$. Patterns vary for model and representation types, but long phrases are generally represented more compositionally.

5.2.1 Named Entities

We used SpaCy to tag and examine named entities (Honnibal and Montani, 2017), as they are expected to be less compositional. We find that named entities indeed have a lower compositionality score in all cases except RoBERTa$_{\text{CLS}}$, indicating that they are correctly represented as less compositional. A representative example is shown in Figure 3. Full results can be found in Appendix J. We break down the compositionality scores of named entities by type and find surprising variation within categories of named entities. For numerical examples, this often depends on the unit used. For example, in RoBERTa$_{\text{AVG}}$ representations, numbers with "million" and "billion" are grouped together as compositional, whereas numbers with quantifiers ("about", "more than", "some") are grouped together as not compositional. The compositionality score distributions for types of named entities are presented in Figure 4.

5.2.2 Examining Compositionality and Phrase Length

There is no consistent relationship between phrase length and compositionality score across models and representation types. However, **CLS** and **AVG** representations show divergent trends. There is a strong positive correlation between phrase length and compositionality score in the **AVG** representations, while no consistent trend exists for the **CLS** representations. This indicates that longer phrases are better approximated as an affine transformation of their subphrase representations. This trend is summarized in Appendix D. All correlations are highly significant.

Phrase type	Most compositional	Least compositional
PP → IN NP	("of", "two perilous day spent among the planters of Attakapas, . . .)	("of", "September")
	("of", "the cloth bandoleers that marked the upper part of his body . . .)	("like", "the Standard & Poor 's 500")
S → NP-SBJ VP	("him", "to suggest it's the difference between the 'breakup' value . . .)	("other things", "being more equal")
	("it", "was doing a brisk business in computer power-surge protectors . . .")	("less", "is more")
NP → NNP NNP	("M.", "Bluthenzweig")	("Edward", "Thompson")
	("Dr.", "Volgelstein")	("Alexander", "Hamilton")

Table 2: Phrases rated most and least compositional using $\text{RoBERTa}_{\text{CLS}}$ representations, from several syntactic phrase types. ". . ." indicates that a phrase continues but is too long to display. Long phrases and abbreviated names tend to have a higher compositionality score.

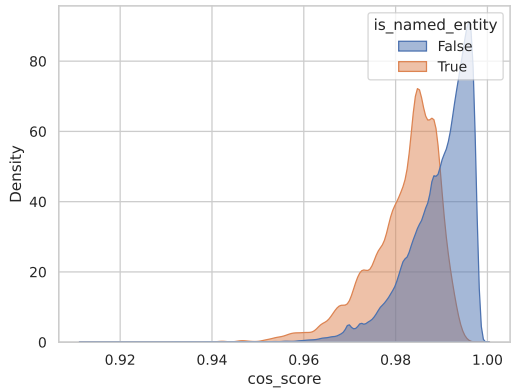

Figure 3: Density plot for compositionality scores of named entities and non-named-entities with $\text{RoBERTa}_{\text{AVG}}$ representations. Higher means more compositional.

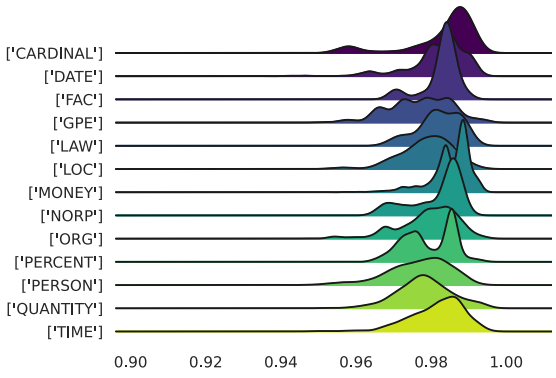

Figure 4: Density plots for compositionality scores of different named entity types with $\text{RoBERTa}_{\text{AVG}}$ representations. Higher means more compositional.

6 Comparing Compositionality Judgments of Humans and Models

6.1 Methods

6.1.1 Human Annotation

Human annotators assigned labels to each phrase in the matched dataset from subsection 3.2: 1 for not compositional, 2 for somewhat compositional, and 3 for fully compositional. They could also decline to answer if they felt that the phrase didn't make sense on its own. Furthermore, they were asked how much each subphrase (left and right) contributed to the final meaning, from 1 for not at all, to 3 for a great deal. The Likert scale of 1-3 was chosen based on analysis of previous compositionality annotation tasks, which found that extreme values of compositionality were the most reliable (Ramisch et al., 2016a).

Initially, six English-speaking graduate students were recruited. The six initial annotators all annotated the first 101 examples and the subset of three annotators with the highest agreement who agreed to continue (Krippendorff $\alpha = 0.5750$) were recruited for the full study, annotating 1001 examples. For the full study, the agreement was higher ($\alpha = 0.6633$). We took the mean of compositionality judgments to be the final score for phrases. The instructions shown to annotators are in Appendix F. Examples judgments from an annotator can be found in Table 3.

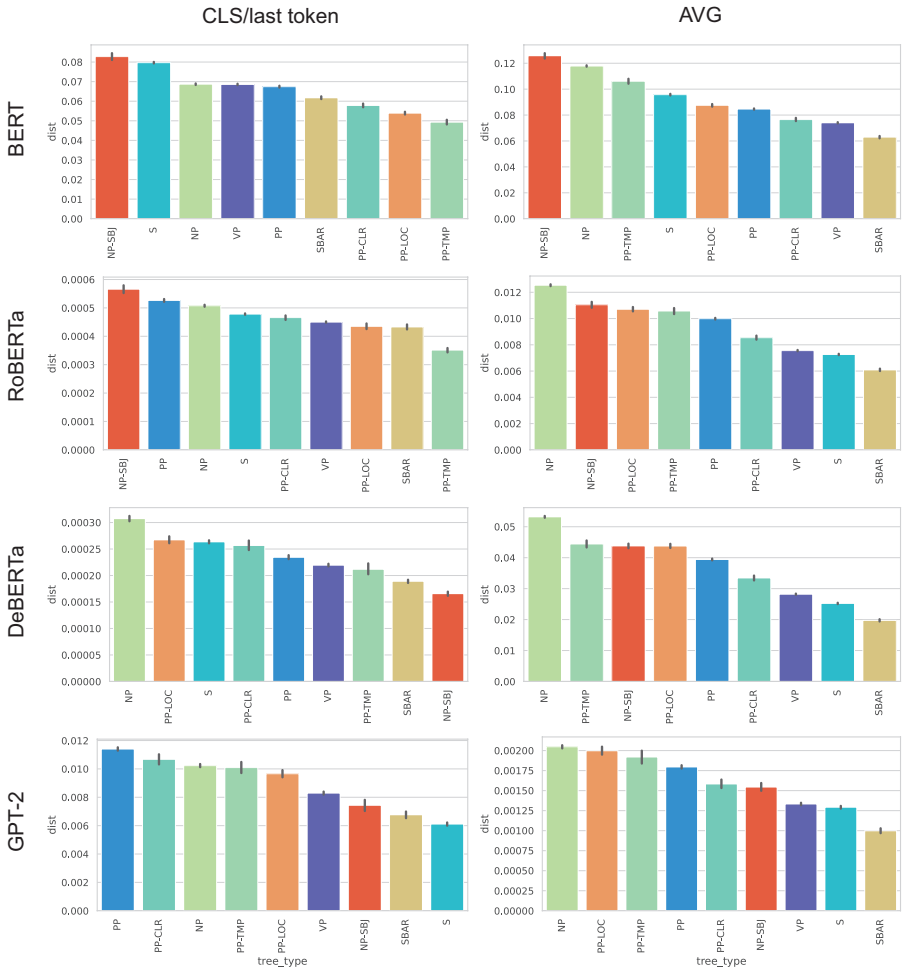

Figure 5: Tree reconstruction error (cosine distance) for each parent phrase type present in the Treebank, ordered from highest mean error to lowest. Based on the affine approximation for each model and representation type. Expanded version with all tree types is presented in Appendix I.

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"	"gravy train"
	"few weeks away"	"shrinking violet"
	"railroad monopoly"	"revolving door syndrome"
$\text{RoBERTa}_{\text{CLS}}$	"two small sticks"	"worse than none"
	"dark glass bottle"	"cases apart"
	"annual music festival"	"arch'd eyebrow"
$\text{RoBERTa}_{\text{AVG}}$	"look with open eyes"	"advertisement revenue"
	"be of equal importance"	"taking it upon oneself"
	"come after breakfast"	"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.

Limitations

One limitation of this work is that it was conducted on a relatively small set of language models trained on English, and the diversity of patterns within even this set of language models and representation types is great. However, we note that the experiments can be easily repeated for any language that has a treebank or good-quality syntactic parsers. A related limitation is that these analyses are dependent on what we take to be the "child" constituents of a parent phrase. It may be harder to examine compositionality for languages that differ substantially from English, or that cannot be easily parsed using existing tools.

Although we try to carefully catalog behaviour observed on natural language phrases, it is likely that smaller-scale experiments providing a more mechanistic understanding of model behaviour would be easier to parse for readers. Although this would be ideal, we leave this for future work, as our main goal was to examine how language models represent phrases considered to be compositional and non-compositional in natural language.

Another limitation is that although we diagnose a problem in language models, we do not provide a clear avenue to fix it. Further work could be done to understand what data distributions or training methods encourage model representations to be more aligned with human judgments. Additionally, although compositionality is linguistically important, more effort could be put towards understanding the downstream tasks for which it is more important. For instance, there could be clear issues in machine translation if non-compositional phrases are not represented properly, but these phrases may not be important in other areas such as instruction following or code generation.

Ethics Statement

Potential Risks and Impacts

Although we aim to document compositionality effects in English, we acknowledge that this perpetuates the problem of English being the dominant language in NLP research. It is possible that conclusions here do not hold for other languages, and further work is needed to understand whether these conclusions transfer.

Additionally, although we tried to filter out offensive idioms from **CHIP**, this was based on one person's best judgment, and it is possible that some of the terms in the dataset may be offensive to some people. Overall, phrases in the dataset tend to be benign, but some idioms are meant to have a perjorative meaning.

Computational Infrastructure and Computing Budget

To run our computational experiments, we made use of a shared compute cluster. We used approximately 100 GPU hours to run experiments, mainly due to running results for different language models and representation types. We did not have any computational budget besides that already used to maintain the cluster.

References

Jacob Andreas. 2019. Measuring compositionality in representation learning. In *7th International Conference on Learning Representations*. ICLR.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Noam Chomsky. 1959. On certain formal properties of grammars. *Information and Control*, 2(2):137–167.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. NeuRIPS.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

M Cresswell. 1973. *Logics and Languages*. Methuen.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022a. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

J Fodor and E LePore. 1992. *Holism: A Shopper's Guide*. Blackwell.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Aurelie Herbelot. 2020. How to stop worrying about compositionality. *The Gradient*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5065–5069. International Joint Conferences on Artificial Intelligence Organization. Journal track.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations*. ICLR.

Brenden M. Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *ArXiv*, abs/1711.00350.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank.

R. McCoy, T. Linzen, E. Dunbar, and P. Smolensky. 2020. Tensor product decomposition networks: Uncovering representations of structure learned by neural networks. In *Proceedings of the Society for Computation in Linguistics*, pages 474–475.

R. Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. 2019. Rnns implicitly implement tensor product representations. In *7th International Conference on Learning Representations*. ICLR.

J Mitchell and M Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, pages 1388–1429.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Barbara H. Partee. 1984. Compositionality. *Varieties of Formal Semantics*, 3:281–311.

Francis Jeffry Pelletier. 1994. The principle of semantic compositionality. *Topoi*, 13:11–24.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Carlos Ramisch, Silvio Cordeiro, and Aline Villavicencio. 2016a. Filtering and measuring the intrinsic quality of human compositionality judgments. pages 32–37.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016b. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Jacob Russin, Roland Fernandez, Hamid Palangi, Eric Rosen, Nebojsa Jojic, Paul Smolensky, and Jianfeng Gao. 2021. Compositional processing emerges in neural networks solving math problems. *CogSci ... Annual Conference of the Cognitive Science Society. Cognitive Science Society (U.S.). Conference*, 2021:1767–1773.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Paul Smolensky, R. Thomas McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. 2022. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems.

Paul Soulos, R. Thomas McCoy, Tal Linzen, and Paul Smolensky. 2020. Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 238–254, Online. Association for Computational Linguistics.

Zoltán Gendler Szabó. 2020. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2020 edition. Metaphysics Research Lab, Stanford University.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

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

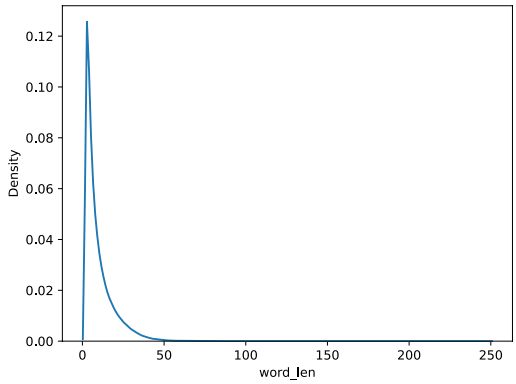

Figure 6: Length distribution of phrases mined from the treebank, in number of words. The modal length was 3 words, followed closely by 2 words. Few phrases contained more than 50 words.

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.

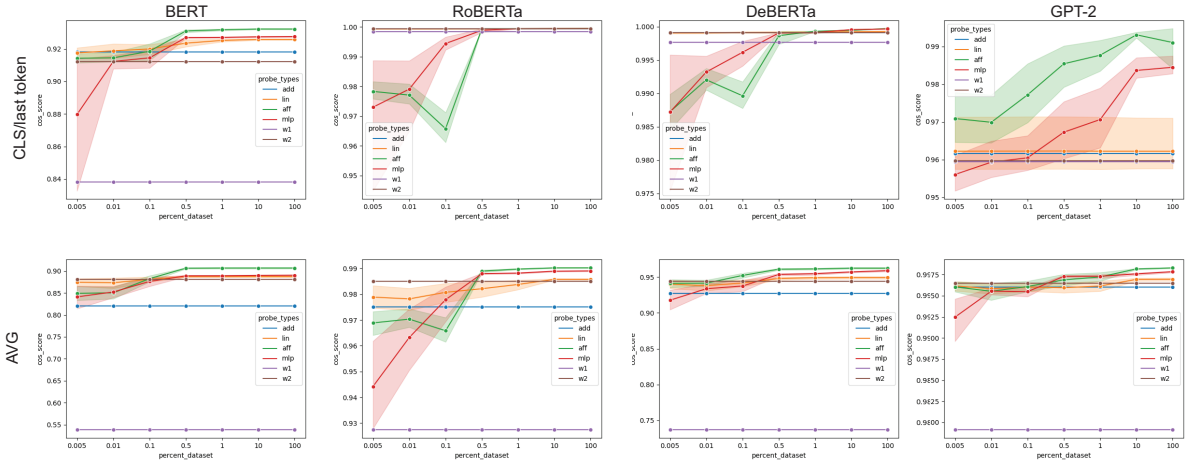

Figure 7: Learning curves of approximative probes trained on differing percentages of train data.

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.

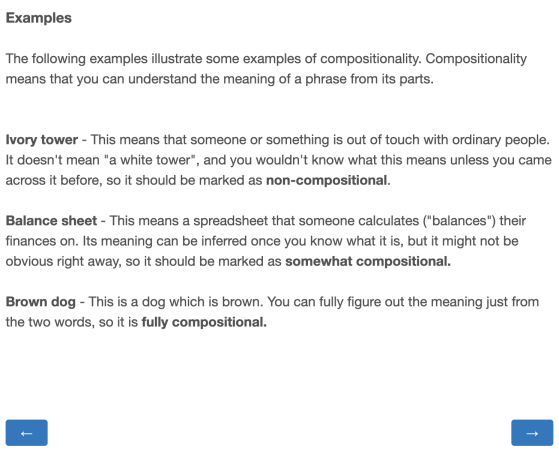

Figure 8: Examples of compositionality judgments shown to annotators

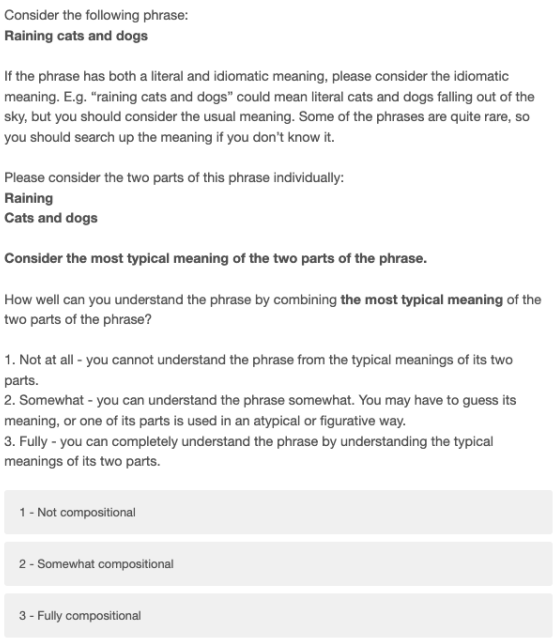

Figure 9: First page of annotation interface for a practice phrase

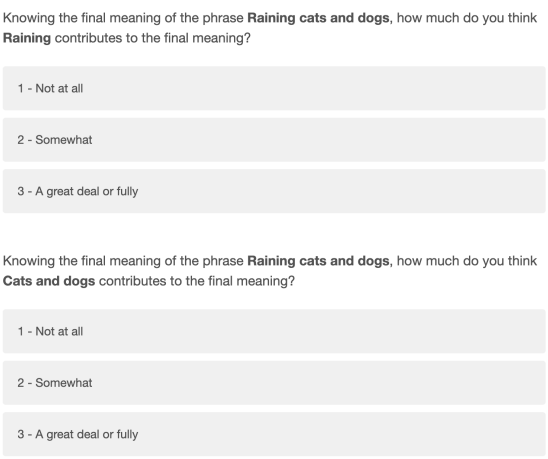

Figure 10: Second page of annotation interface for a practice phrase

G Compositionality scores without anisotropy correction

The raw compositionality scores can be found in Table 10.

H AUC of approximative probes

Model and representation	probe	AUC
$BERT_{CLS}$	ADD	91.80
	W1	83.82
	W2	91.20
	LIN	92.57
	AFF	**93.20**
	MLP	92.74
$RoBERTa_{CLS}$	ADD	99.93
	W1	99.84
	W2	99.93
	LIN	99.94
	AFF	99.93
	MLP	**99.94**
$DeBERTa_{CLS}$	ADD	99.90
	W1	99.75
	W2	99.90
	LIN	99.92
	AFF	99.94
	MLP	**99.95**
$GPT\text{-}2_{last}$	**MLP**	**99.95**
	ADD	96.16
	W1	95.94
	W2	95.97
	LIN	96.21
	AFF	**99.18**
	MLP	98.32
$BERT_{AVG}$	ADD	82.04
	W1	53.83
	W2	88.10
	LIN	88.68
	AFF	**90.63**
	MLP	88.96
$RoBERTa_{AVG}$	ADD	97.51
	W1	92.73
	W2	98.49
	LIN	98.56
	AFF	**99.00**
	MLP	98.88
$DeBERTa_{AVG}$	ADD	92.74
	W1	73.67
	W2	94.38
	LIN	94.89
	AFF	**96.21**
	MLP	95.75
$GPT\text{-}2_{AVG}$	ADD	99.60
	W1	97.90
	W2	99.64
	LIN	99.69
	AFF	**99.81**
	MLP	99.76

Table 11: AUC scores for probes trained on various percentages of the training set.

I Mean deviation of phrase types by tree type

The mean deviation of the most common tree types can be found in Figure 11.

J Further named entity results

Named entity results can be found in Figure 12 and Figure 13.

Model and representation	Probe	Mean reconstruction score	Standard dev.
BERT_{CLS}	ADD	0.9178	0.001159
	W1	0.8382	0.003599
	W2	0.9117	0.0007133
	LIN	0.9258	0.0002285
	AFF	**0.9322**	0.0002033
	MLP	0.9276	0.0002108
$\text{RoBERTa}_{\text{CLS}}$	ADD	0.99935	3.895×10^{-6}
	W1	0.99850	2.612×10^{-5}
	W2	0.99937	6.866×10^{-6}
	LIN	0.99946	4.735×10^{-6}
	AFF	**0.99950**	6.093×10^{-6}
	MLP	0.99947	4.719×10^{-6}
$\text{DeBERTa}_{\text{CLS}}$	ADD	0.99908	4.070×10^{-5}
	W1	0.99762	2.900×10^{-5}
	W2	0.99911	1.399×10^{-4}
	LIN	0.99928	8.963×10^{-5}
	AFF	**0.99972**	1.542×10^{-5}
	MLP	0.99965	2.323×10^{-5}
BERT_{AVG}	ADD	0.8205	0.0003836
	W1	0.5383	0.007471
	W2	0.8893	0.03071
	LIN	0.8873	0.003071
	AFF	**0.9069**	0.002566
	MLP	0.8904	0.002988
$\text{RoBERTa}_{\text{AVG}}$	ADD	0.9752	0.0001306
	W1	0.9274	0.001695
	W2	0.9850	0.0005092
	LIN	0.9858	0.0004573
	AFF	**0.9902**	0.0003076
	MLP	0.9890	0.0003981
$\text{DeBERTa}_{\text{AVG}}$	ADD	0.9275	0.002634
	W1	0.7368	0.001575
	W2	0.9438	0.003321
	LIN	0.9493	0.003036
	AFF	**0.9625**	0.001814
	MLP	0.9590	0.002145
$\text{GPT-2}_{\text{AVG}}$	ADD	0.9960	0.0002833
	W1	0.9791	0.0001214
	W2	0.9965	0.0003359
	LIN	0.9970	0.0003036
	AFF	**0.9984**	0.0002617
	MLP	0.9979	0.0001634

Table 10: Mean reconstruction score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Not corrected for anisotropy in each representation/model type.

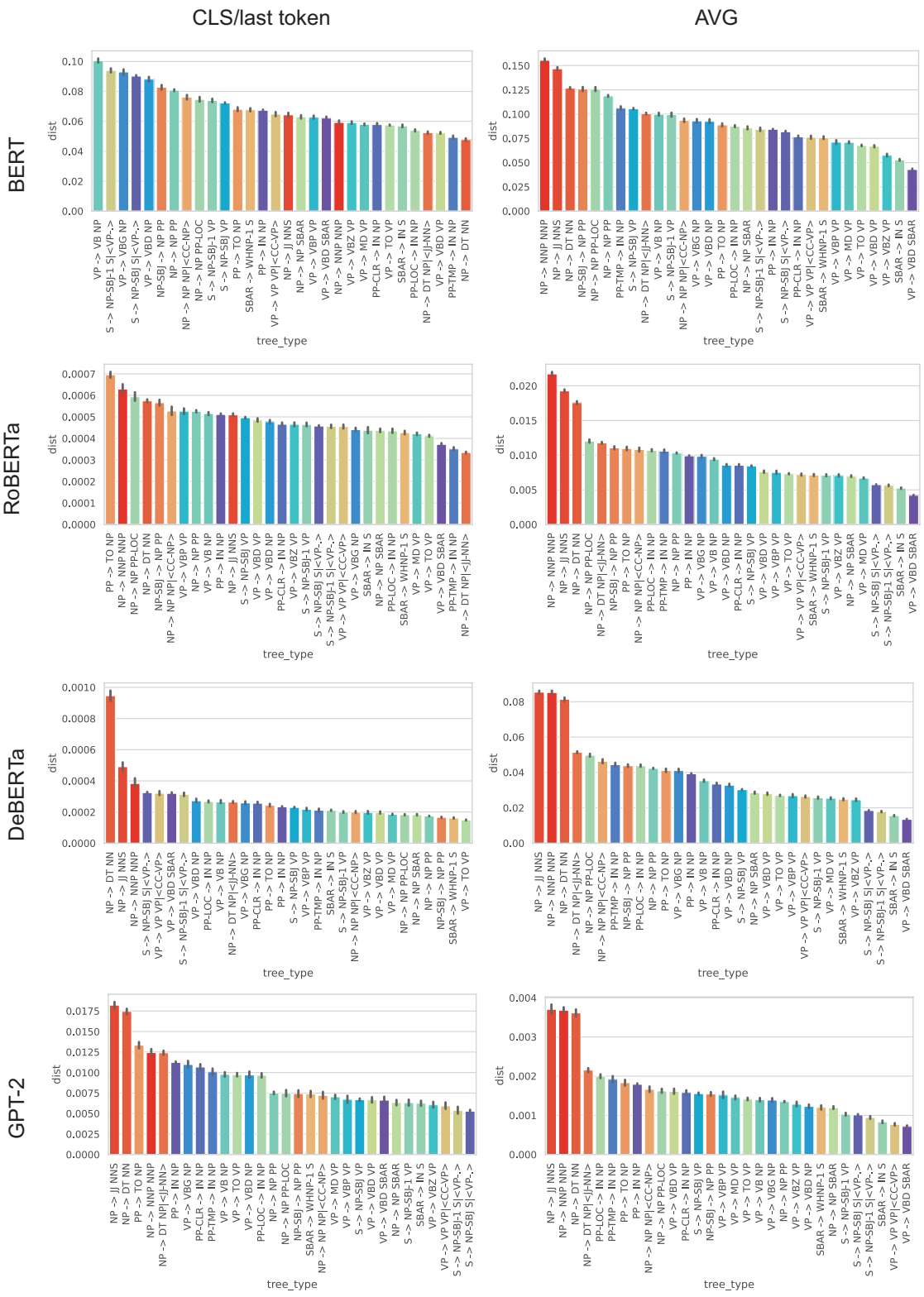

Figure 11: Mean deviation from predicted representation across full tree types.

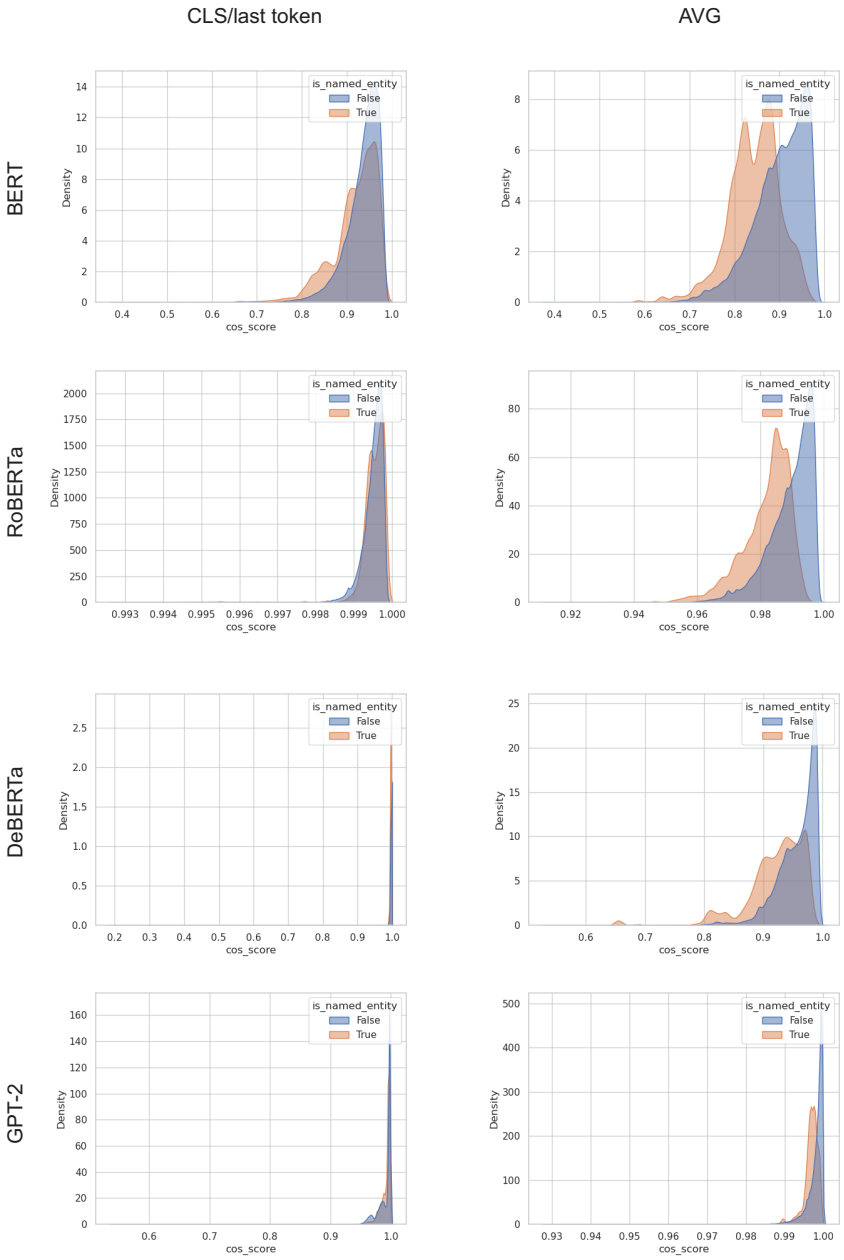

Figure 12: Distributions of compositionality score for named entities and non-named entities across model types and representation types. The AVG representation matches the intuition that named entities are usually less semantically compositional, as they point to an entity in the real world that may not relate to their name.

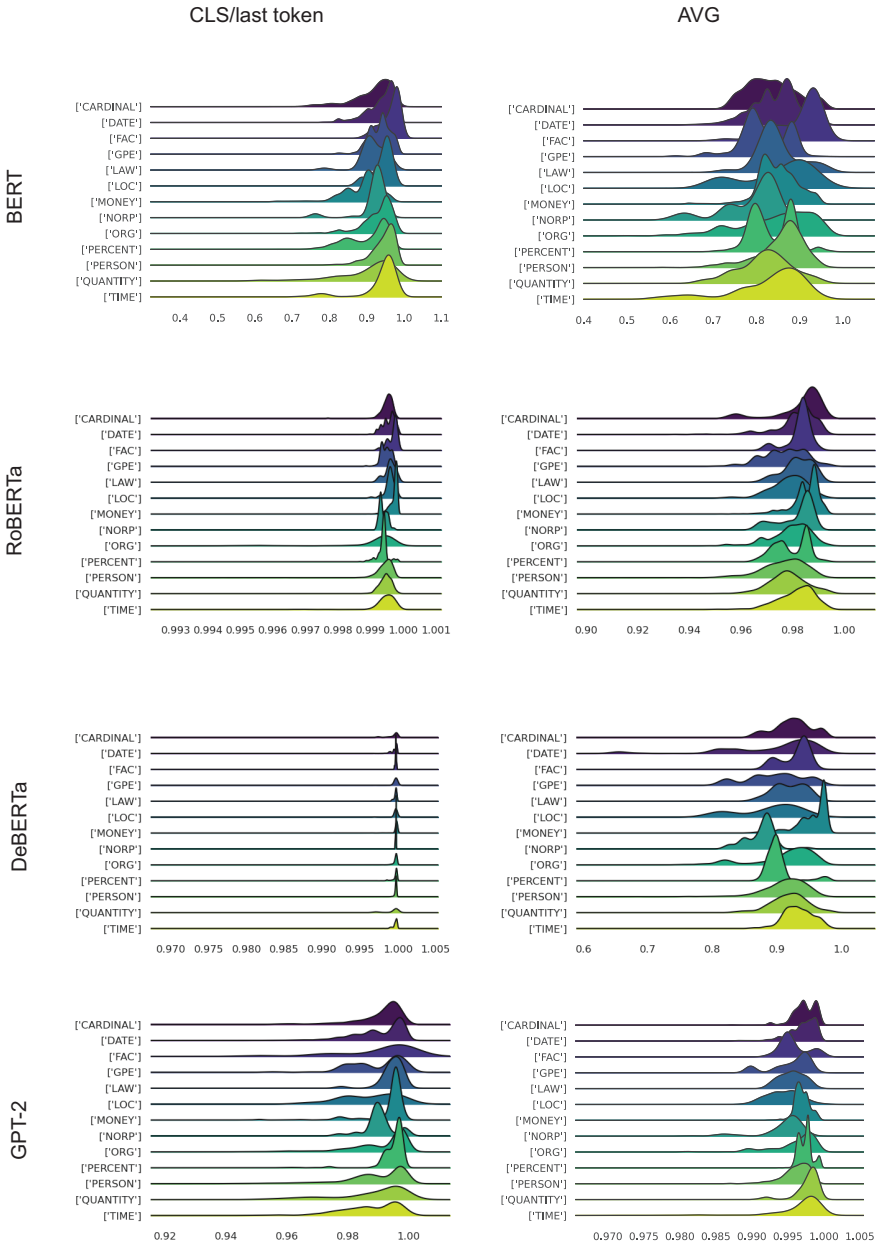

Figure 13: Visualization of distribution of compositionality scores across different types of named entities.

K Frequency and length correlations

Model and representation	Feature	Spearman ρ	p-val
BERT_{CLS}	Word length	0.2182	3.055×10^{-10}*
BERT_{AVG}		0.007396	0.08722
$\text{RoBERTa}_{\text{CLS}}$		0.01686	0.6193
$\text{RoBERTa}_{\text{AVG}}$		0.3653	4.773×10^{-28}*
$\text{DeBERTa}_{\text{CLS}}$		0.4087	1.709×10^{-35}*
$\text{DeBERTa}_{\text{AVG}}$		0.4484	1.340×10^{-42}*
$\text{GPT-2}_{\text{last}}$		0.3228	8.481×10^{-22}*
$\text{GPT-2}_{\text{AVG}}$		0.0.3125	1.719×10^{-20}*
Human	Word length	0.05666	0.1894
BERT_{CLS}	Frequency	0.2182	0.08193
BERT_{AVG}		-0.08582	0.07899
$\text{RoBERTa}_{\text{CLS}}$		0.02548	0.9053
$\text{RoBERTa}_{\text{AVG}}$		-0.08354	0.08193
$\text{DeBERTa}_{\text{CLS}}$		-0.1265	0.001459*
$\text{DeBERTa}_{\text{AVG}}$		-0.2185	6.455×10^{-10}*
$\text{GPT-2}_{\text{last}}$		-0.05750	0.3595
$\text{GPT-2}_{\text{AVG}}$		0.04382	0.5891
Human	Frequency	0.008363	0.9053

Table 12: Correlations of frequency and length with human and model compositionality scores. Corrected with Holm-Bonferroni correction.