Language use is only sparsely compositional: The case of English adjective-noun phrases in humans and large language models

Aalok Sathe[1] **(asathe@mit.edu)**
Evelina Fedorenko[1,*] **(evelina9@mit.edu)**
Noga Zaslavsky[2,3,*] **(nogaz@uci.edu)**

[1]Massachusetts Institute of Technology, [2]UC Irvine, [3]New York University

Abstract

Compositionality is considered a key hallmark of human language. However, most research focuses on item-level compositionality, e.g., to what extent the meanings of phrases are composed of the meanings of their sub-parts, rather than on language-level compositionality, which is the degree to which possible combinations are utilized in practice during language use. Here, we propose a novel way to quantify the degree of language-level compositionality and apply it in the case of English adjective-noun combinations. Using corpus analyses, large language models, and human acceptability ratings, we find that (1) English only sparsely utilizes the compositional potential of adjective–noun combinations; and (2) LLMs struggle to predict human acceptability judgments of rare combinations. Taken together, our findings shed new light on the role of compositionality in language and highlight a challenging area for further improving LLMs.

Keywords: language; compositionality; semantics; large language models; information theory

		Counts
COCA	Adjectives	147K
	Nouns	469K
	Adj-N combinations (total)	25M
	Adj-N combinations (unique)	4.4M
Baselines	Theoretically possible	69B
	MaxComb	8.9M ($\pm$1.75K)

Table 1: Corpus statistics from COCA, in comparison with two baselines: all theoretically possible Adj-N combinations, and combinations generated by independently sampling 25M Adj-N pairs based on their marginal corpus frequencies (MaxComb, empirical STD is over 100 repetition). Observed COCA combinations occupy only a tiny fraction of all theoretically possible combinations, and roughly half of the independently sampled combinations.

Introduction

Compositionality extends the information language can encode by allowing units to combine with one another in new ways and produce novel meanings (Partee et al., 1984). Compositionality is seen as one of the core principles that allowed human communication systems to flourish (Frankland & Greene, 2020; Johnson, 2020; Smith & Kirby, 2012; Fodor & Pylyshyn, 1988). Experiments with emergent languages show compositionality emerging under efficiency constraints (Chaabouni et al., 2020). However, the extent to which compositionality prevails in human languages remains unclear: most prior literature and research focuses on phrase-level compositionality (e.g., Arnon & Snider, 2010; Morgan & Levy, 2016), that is, to what extent the meanings of phrases are composed of the meanings of their sub-parts, rather than on system-level compositionality, i.e., the degree to which a linguistic system as a whole utilizes the space of possible item combinations. Here, in a series of three studies, we explore the extent to which compositionality is utilized in language, focusing on the case of English adjective-noun (Adj-N) combinations as a testbed.

First, we propose a new information-theoretic measure for system-level compositionality and apply it to the space of English Adj-N combinations using corpus analyses and LLM probabilities to ask how much of the space of possible linguistic combinations language use actually covers (Study 1). We find that the vast majority of two-word Adjective-Noun phrases are practically unrealized even in a large corpus (the Corpus of Contemporary American English; COCA; Davies, 2009), and the distribution of occurrences of two-word phrases is skewed above and beyond a simple baseline that maximally spans the set of combinations respecting the marginal Zipfian distribution of words. Furthermore, we find that LLMs predict even more sparsity in this compositional semantic space compared to COCA after regularization with marginal lexical priors from COCA (Figure 1). These results suggest that out of a vast number of possible (syntactically allowed) word combinations, only a tiny fraction appears to be useful and/or actually used in natural language. Second, to control for finite-sample effects, we collected new human acceptability judgments for rare combinations (Study 2), confirming that they are indeed mostly non-sensible.

Finally, given the remarkable success of LLMs, we ask whether humans and LLMs align in their judgments of these rarely-observed Adj-N combinations (Study 3). Although recent LLMs have been shown to capture linguistic meaning well enough to perform diverse linguistic tasks (Liang et al., 2022; Pavlick, 2022), such tasks typically involve word combinations that occur in typical language use. There are concerns that LLMs find it difficult to generalize to meaning-spaces and tasks not sufficiently represented in their training

* Equal senior contribution.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

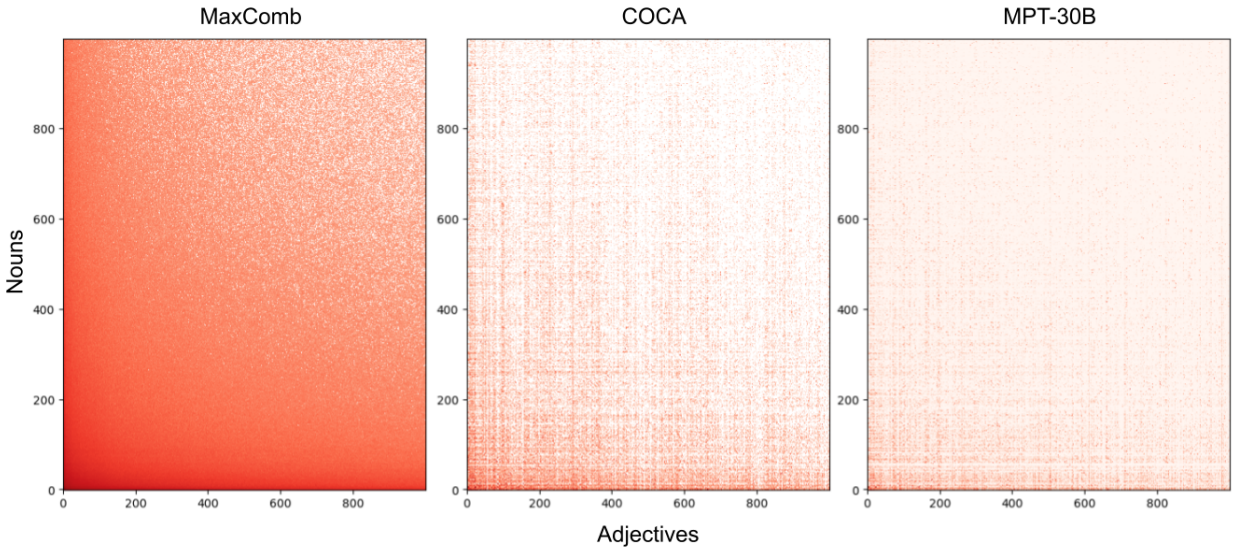

Figure 1: Normalized frequencies of Adj-N pairs for the 1000 most frequent adjectives and nouns in COCA, based on (a) the MaxComb baseline, (b) COCA counts, and (c) MPT-30B probabilities (adjusted to match the COCA marginal adjective distribution). Adjectives and nouns are sorted by their corpus frequencies in COCA. Both COCA and MPT-30B reflect substantially lower degree of compositionality compared to the MaxComb baseline, which reflects the maximal degree of compositionality afforded by respecting the marginal frequencies of adjective and nouns.

data (Kauf et al., 2023). How LLMs might judge the compositional space of word combinations most of which are unattested in a large corpus and the rest rarely used is an open question. To stress-test the similarity of the underlying compositional spaces, we focused on combinations that are unattested in English and asked whether LLMs partition the space into "meaningful" versus "nonsensical" components in a way similar to humans. To do so, we conducted an online experiment and collected behavioral sensibility judgments ("How much sense does this make?" on a 1 to 7 scale) from $n = 1000$ participants on roughly $10,000$ unattested adjective-noun combinations (sampled from the stimuli of Vecchi et al. (2017)), resulting in ≥ 8 judgments per adjective-noun pair. Human judgments show good inter-rater consistency, suggesting that they capture meaningful variation in the compositional space, even for out-of-distribution linguistic combinations. We find that LLMs struggle to predict human judgments on these very rarely seen, stimuli, although they do capture the underutilized compositionality in adjective-nouns pairs that was also observed in the COCA corpus. This demonstrates a gap between human compositional semantics and that modeled by LLMs and suggests a new challenging benchmark for further improving LLMs in future work.

Related Work

Compositionality in language Christiansen & Chater (2015) observe that though language allows for a high degree of combinatorial composition, there is a disparity between what is possible and what is easy to comprehend and produce. The authors observe that people rarely produce, and have trouble understanding, multiply-embedded recursive structures. In the context of Adjective-Noun composition, whereas we have a very large space of possible combinations, it may be the case that people find it easier to use more frequently encountered and easily composed combinations. Our study helps lay the first step towards this question: how compositional is the space?

Computational models of compositionality Lapata et al. (1999) collected human judgments on 90 attested Adjective-Noun pairs (30 adjectives and 1 noun each from low-, medium-, and high-frequency buckets) and used corpus-based metrics to predict the acceptability. The authors found a high correlation between humans' ratings and corpus frequency of Adj-N pairs. Biemann & Giesbrecht (2011) describe a workshop shared task on distributional semantics and compositionality that considered Adj-N as well as other multi-word expression compositionality. The authors report most distributional semantic features used in contributions to the workshop did better than frequency baselines yet struggled to mirror human ratings. In a previous study, Vecchi et al. (2017) collected people's preferences of Adjective-Noun pairs they thought made more sense in a binary forced-choice paradigm. The authors presented participants with two Adj-N pairs side-by-side in each trial and asked them to pick the one that made more sense. The study used a total of $27k$ AN pairs and presented them as paired items (AN_i, AN_j). The study then compiled these preferences into a score per Adj-N pair by counting the number of times the pair was in a pref-

erential position, of all the times the pair occurred as a trial. The authors use multiple measures ranging from the surface form of the items to their compositional vector-space representations proposed by previous literature to try to provide a computational account of their ratings.

In the backdrop of past attempts to model compositional understanding of short phrases, it is an open question whether current computational models have more nuanced compositional semantics. More recently, Large Language Models (LLMs) make a compelling case for understanding varied language use with their high performance on diverse language tasks. GPT2-XL, the smallest model in our set, has been shown to be effective in accurately predicting brain responses and behavioral measures to diverse linguistic stimuli (Hosseini et al., 2022; Tuckute et al., 2024). Similarly, Liu et al. (2022) demonstrate LLMs' ability to understand figurative language using metaphorical expressions. Considering these feats, it is a natural question whether LLMs will capture the structure of the compositional space of Adjectives and Nouns in English, and whether their assessments align with those of humans.

Study 1: Measuring the degree of utilized compositionality

Methods We employ two approaches for estimating word-pair frequencies, which is the basis for our quantitative measure of compositionality. First, we use the Corpus of Contemporary American English (COCA; Davies, 2009) (containing 385M+ words) as a view of observed language use across the years 1991-2012. We use 'stanza' (Qi et al., 2020) to dependency-parse the corpus and universal parts-of-speech (UPOS) to identify adjectives and nouns. For our analyses, we picked all the Adjective-Noun pairs that were in an *amod* dependency relation, with the noun in a *parent* and adjective in a *child* relationship to one another. Table 1 shows the counts of individual lexical items of each category. Naively combining all lexical items with one another would lead to 69B Adjective-Noun pairs. We observe 4.4M pairs with at least one occurrence in the corpus. Given the finite nature of the corpus, it is unreasonable to expect 69B combinations to be realized. Instead, we can set an expectation based on the marginal distributions of Adjectives and Nouns treating them as independent random variables. We sample as many Adjective-Noun pairs as are observed in the corpus: 25M. Each pair is constructed by sampling an Adjective and Noun from their respective marginal lexical distributions in the corpus with repetition. We call this baseline **MaxComb**, because it will generate the maximal number of unique combinations under a constraint on the marginal distribution of adjectives and nouns. By repeating this procedure k times, we derive an empirical uncertainty estimate for this baseline.

Second, we use large language models (LLMs) as models of the compositional semantic space. Whereas COCA is a large corpus spanning decades of language use, LLMs are trained on much larger datasets from varied sources, including data from the internet, and as such, offer a different view into language use. To make our analyses and experiments with LLMs tractable, for the remainder of this paper we restrict ourselves to a subset of the 1000 most frequent Adjectives and Nouns as observed in COCA, and the 10^6 (1M) theoretically possible combinations that can result. We construct a joint distribution by normalizing over the total occurrences over these 1M Adj-N pairs. We observe a total of 11.7M and $336,473$ (0.3M) unique combinations within this subspace in COCA. Next, we estimate LLM probabilities over sequences using the setup "$P_{LLM}(n_j \mid \textit{How likely is this: } a_i)$" (written in shorthand hereafter as $P_{LLM}(N|A)$). Upon testing additional prompts/contexts we find high correlation in the probability estimates produced by using different prompts and that the choice of prompt plays little role. We use several autoregressive models of sizes varying from 1.5B parameters to 30B parameters. We use models off-the-shelf that are pre-trained and used without any modification. For some models, where available, we include their chat-optimized variants. The models we used were: GPT2-XL (1.5B), Phi-2 (2.7B), Mistral (7B), MPT (7B, 30B, 7B-chat, 30B-chat). To measure similarity with corpus-based probabilities $P_{COCA}(N|A)$, we condition the joint distribution on Adjectives and calculate the Pearson's R correlation coefficient.

To quantify the extent to which compositionality is utilized, we turn to information theory. We take the joint entropy of adjectives and nouns, $H(A,N)$, as a measure of their utilized compositionality. Intuitively, if all Adj-N pairs are equally utilized, that will result in maximal joint entropy. On the other hand, if only a single (a,n) pair is utilized, that will yield $H(A,N) = 0$. To account for other constraints in language, we assume that the marginal distributions of adjectives and nouns, i.e., $p(A)$ and $p(N)$ respectively, are fixed. We can then ask, given this word-frequency constraint, what is the degree of unutilized compositionality in the empirically observed joint distribution $p(A,N)$? Using the following well-known identity that relates entropy with mutual information (Cover, 1999):

$$H(A,N) = H(A) + H(N) - I(A;N)$$

one can see that if the marginal distributions are known and fixed, then the corresponding marginal entropies $H(A)$ and $H(N)$ are also fixed, and $I(A;N)$ captures the extent to which the joint distribution reflects compositional structure: $H(A,N)$ would be maximal when $I(A;N) = 0$, in which case adjectives and nouns are combined by independent sampling (as in the MaxComb baseline), yielding maximal utilized compositionality given fixed marginals. $H(A,N)$ would be minimal when the mutual information between adjective and nouns is maximal, for example, when each noun can be combined with only a single adjective. Therefore, we take $I(A;N)$ as our measure for unutilized compositionality.

Results Our sampling procedure leads to 8.9M unique pairs ($SD = 1750$ from $k = 100$ repetitions), about twice as many as

	$H(A)$	$H(N)$	$H(A;N)$	$I(A;N)$
COCA	8.476	9.263	15.327	2.413
MaxComb	8.476	9.263	17.740	0.000
MPT-30B	8.476	9.019	14.923	2.573

Table 2: We calculate the marginal and joint entropies (bits) of the joint Adjective-Noun distributions for the 1000 most frequent lexical items. We consider three sources: Adj-N combinations observed in COCA; the result of a random baseline (MaxComb); and those predicted by conditional probabilities from MPT-30B adjusted with the corpus marginal Adjective distribution.

the unique pairs observed in COCA resulting from the same total number of combinations (25M), described in Table 1. Table 2 outlines the mutual information of the joint distribution of Adjectives and Nouns, computed for each of the three different sources of estimating compositionality. Entropies for joint and marginal distributions can be empirically determined for joint distributions that are fully observable. **COCA** corresponds to the observed distribution, restricted to the top 1000 items, of Adj-N in the corpus. **MaxComb** corresponds the the joint distribution of marginal Adjective and Noun distributions from COCA. **MPT-30B** corresponds to conditional probabilities $P_{MPT\text{-}30B}(N|A)$ adjusted against the marginal distribution $P_{COCA}(A)$ of Adjectives to get a joint distribution over Adj-Ns. The resultant three joint distributions all share the marginal Adjective distribution. However, the marginal Noun distribution of MPT-30B differs. We observe $I_{MPT\text{-}30B} > I_{COCA} > I_{MaxComb}$, with the mutual information of the random baseline being 0, as expected (pairs are combined in a non-systematic manner). The results suggest that MPT-30B estimates the sparsity of the compositional space to be even higher than that observed in a finite but large corpus. Figure 1 best helps illustrate this observation. The compositional space of Adjective-Noun combination is under-utilized compared to what their marginal distributions might suggest—people use few combinations with intention rather than many combinations more widely and flexibly.

Figure 2 shows R^2 values for correlations between the corpus conditional probabilities $P_{COCA}(N|A)$ and those from LLMs, $P_{LLM}(N|A)$. We find that the two distributions are somewhat correlated (R^2 not exceeding 0.26), suggesting that though COCA and LLMs both estimate sparsity in language to be high, they may differ in how the sparsity is distributed within this compositional space.

The quantification and modeling of this compositional space raises questions for future work: (1) How can we explain the observed sparsity? Do people re-use adjective-noun combinations even though many more meaningful but infrequently-used combinations exist? (2) How much of the sparsity effect is driven by a finite-sample effect of either the corpus or LLMs' training data? To answer some of these questions, we turn to the next part by collecting judgments

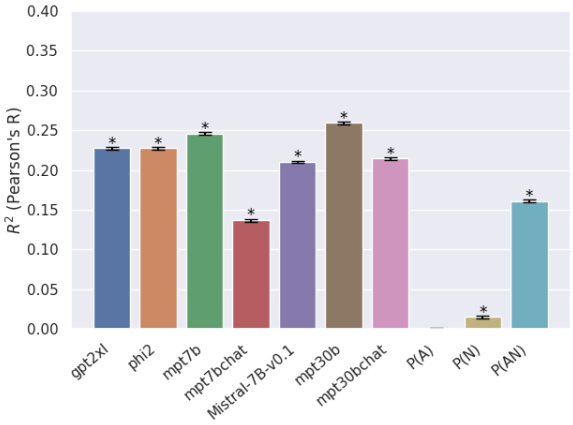

Figure 2: R^2 from Pearson correlation between $P_{COCA}(N|A)$ based on corpus statistics and $P_{LLM}(N|A)$. A star above the bar indicates $p < .001$. Error bars denote 95% confidence intervals for the correlation. In addition to LLMs we show three baselines for comparison with the corpus conditional distribution: the marginal distributions of Adj and N, as well as their joint distribution.

on out-of-distribution data and computationally modeling it.

Study 2: Are unattested Adj-N combinations meaningful?

In this study, we collected sensibility judgments for 10,000 items. Our set of 10,000 is a random subset of the 27,000 stimuli used by Vecchi et al. (2017) in a previous study. Whereas the authors of the original study collected forced-choice judgments and then aggregated them, we use Likert-scale ratings as a way to get more direct per-item ratings with a large number of raters per item, to be able to compare ratings on individual Adj-N pairs more widely outside of the stimulus set.

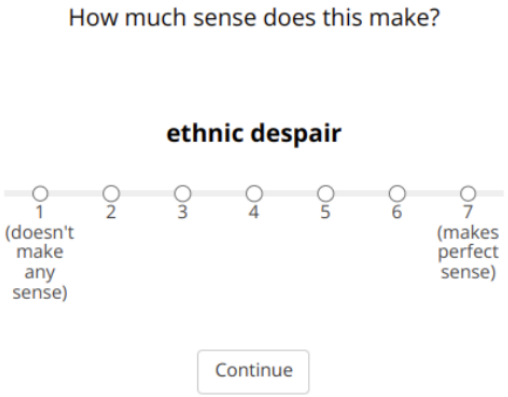

Figure 3: Example of item presentation in the online study.

Methods Participants on Prolific were recruited based on self-reported fluency in English. Of the 1000 participants,

629 identified as male, 370 as female, and 1 preferred not to answer. Participants were each asked to rate 200 adjective-noun pairs based on "whether it makes sense on a scale from 1 (doesn't make any sense) to 7 (makes perfect sense)" (as shown in Fig. 3. As examples, participants were told that pairs like "nice family" or "scary movie" make a lot of sense, but pairs like "educated hairbrush" or "deep-fried ballerina" don't make much sense. Participants were told to judge each pair on its own regardless of whether a pair may become plausible in a larger context. For example, the pair "bisyllabic family" doesn't make much sense on its own even though it could become more sensible if it was followed by another word, like "bisyllabic family name". Each participant saw a random subset of the full set of items, and participants could only do the study once. For most items, we were able to obtain about 10 ratings. In addition to 200 critical items, we included 12 high-collocation 'gold' items attested in a corpus and 18 unattested items for a total of 30 items shared across all participants to compute inter-rater agreement.

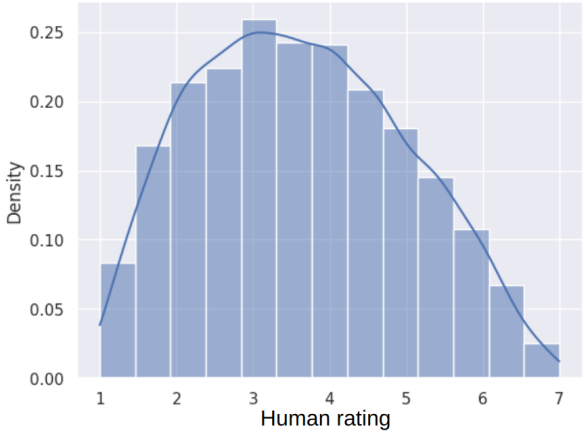

Figure 4: Human acceptability ratings for unattested Adj-N pairs, on a Likert scale from 1 (doesn't make any sense) to 7 (makes perfect sense).

Results We computed split-halves and leave-one-out (LOO) correlations on collected data. The mean split-halves correlation with 1000 bootstrapped iterations was 0.99. The mean LOO correlation was 0.826. Participants with LOO correlation lower than 0.4 or $p \geq .05$ were excluded. After exclusion, we were left with 9907 items with at least 8 ratings each. Table 3 shows examples ranging in their acceptability judged by humans. Using an ordinary least-squares (OLS) regression model we find that our data correlates with the composite score derived from the data collected by Vecchi et al. (2017) for the same items with $R^2 = 0.66$ using Pearson's R.

Figure 4 shows the resultant distribution of ratings from humans, averaged over item. Overall, most pairs receive low ratings, suggesting that compositionality is indeed underutilized. However, there is still a substantial amount of high ratings, suggesting that both corpus frequencies and LLMs underestimate the degree of compositional meaning in actual language use.

A	N	$\hat{P}(A)$	$\hat{P}(N)$	Avg. rating
yellow	certainty	.172	.032	1.00
pregnant	republic	.108	.011	1.00
married	weapon	.001	.095	1.00
entire	surveyor	.384	.003	1.72
native	subscription	.199	.015	1.75
twin	recycling	.034	.035	2.87
amazing	emperor	.155	.022	7.00
lucky	grandmother	.127	.112	7.00

Table 3: Example items from the stimulus set, sampled from 1 (lowest-rated) to 7 (highest-rated). $\hat{P}$ is shorthand for $P_{COCA} \times 10^3$, modified in this manner for readability. Average rating for an item is the average across all participants who provided a rating on the item. All items included in the analysis received at least 8 ratings.

Study 3: Can LLMs predict human acceptability ratings?

LLMs and the corpus provide two different accounts of sparsity in language. Since both are approximations, it is unclear which one is closer to true language use. However, both LLMs and the corpus predict a high amount of sparsity. In this section we ask whether LLMs can mirror human ratings for a set of stimuli that are either unattested or occur only very rarely, and are likely to have different characteristics compared to to the LLMs' training data.

Methods We used conditional probability estimates from LLMs as described previously to evaluate correlation with Likert-scale ratings averaged across participants per item on a scale from 1 to 7 (examples in Table 3). In addition to our set of LLMs, we also use $P(A)$ and $P(N)$, derived from COCA, as baseline metrics to see how well human judgments align with marginal lexical distributions. To evaluate our models we use Spearman's rank-order correlation coefficient and compute the R^2 to measure the similarity between the distribution of human ratings and LLM-estimated conditional probabilities. Spearman's correlation coefficient allows us to compare the spaces of judgments from humans and those from LLMs without needing to align them in a shared judgment- or probability-space.

Results We find that LLMs do poorly in predicting human judgments on this set of items largely unattested in COCA. An as example, Figure 5 shows the distribution of average ratings per item and corresponding LLM-estimated conditional probabilities $P(N|A)$ (log-scaled and normalized) from MPT-30B. Figure 6 shows R^2 derived from Spearman's correlation coefficient. Neither LLMs nor baseline measures exceed

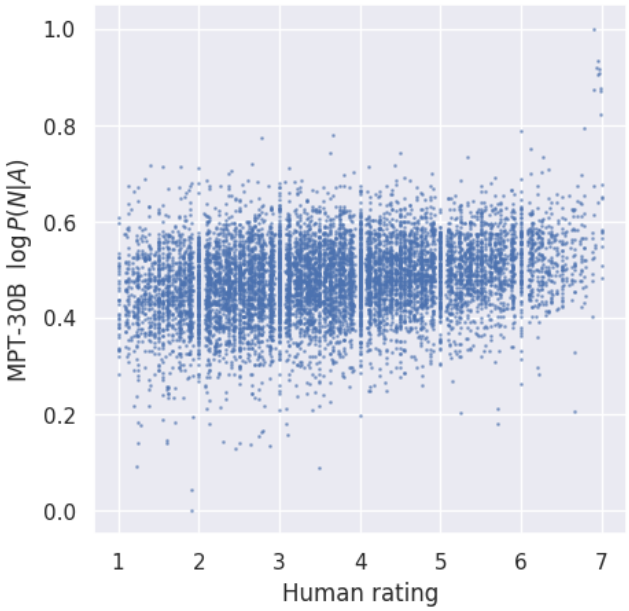

Figure 5: MPT-30B struggles to predict human ratings. Each dot represents an Adj-N pair used in the. The x-axis represents the average rating from humans on an Adj-N pair on a Likert scale from 1 (doesn't make any sense) to 7 (makes perfect sense). The y-axis represents conditional probabilities estimated by MPT-30B (log-scaled and normalized).

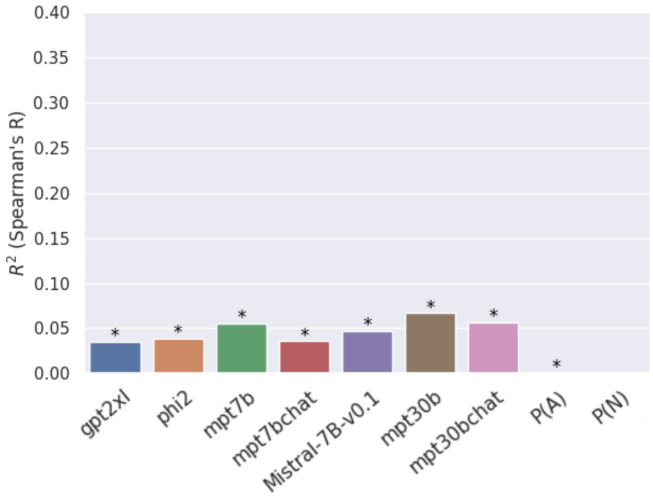

Figure 6: LLMs perform poorly in predicting humans' sensibility judgments on unattested stimuli. The y-axis denotes the R^2 corresponding to Spearman's R computed over average rating per item judged by humans and the conditional probability $P_{LLM}(N|A)$ estimated by a number of LLMs described previously, arranged in the order of number of parameters. As before, $P(A)$ and $P(N)$ denote marginal corpus distributions of Adjectives and Nouns. We use Spearman's R here to compare two non-congruent scales: one, a Likert scale from 1 to 7, and the another, a probability distribution. A star above bars indicates $p < .001$.

$R^2 = 0.1$, indicating poor alignment between acceptability judgments and LLM estimates. The result also highlights that human judgments cannot be predicted by marginal lexical distributions, highlighting the necessity of compositional semantic understanding in being able to judge the Adj-N pairs. Future evaluation could consider whether items more easily agreed-upon across participants lead to better human-LLM alignment.

Discussion

Our work provides a novel quantification of compositionality in English using corpus analyses and empirical data from humans and LLMs. Both, a large corpus and LLMs suggest that the compositional space of Adj-N combinations in English is highly sparse—people seem to make use of a only small fraction of this space. In follow-up work we would like to better understand and model what combinations are more likely and why people choose to use them rather than make wider use of the flexibility. Approaches such as co-clustering the two distributions maximizing mutual information (Dhillon et al., 2003) would yield interpretable clusters that combine highly with one another, allowing us to interpret why the compositional space is used the way it is.

Large language models (LLMs) provide an account of sparsity similar in scale to that seen in corpus data, likely mirroring what LLMs consider in-distribution vs. out. However, when it comes to unattested stimuli, or out-of-distribution compositions, LLMs do not mirror humans' judgments—that humans are highly consistent in—on the sensibility of Adj-N pairs, suggesting a gap in the compositional semantics that can be induced solely from distributional data: formal linguistic competence (Mahowald et al., 2024) does not necessitate rich semantic understanding. Though it is possible, albeit unlikely, that the models we considered in fact are encoding information about human sensibility judgements, but that information is not read out from their predicted probabilities over text. In future work we could also consider using alternate approaches of obtaining this information using linear projection and prompting.

While we focused here on Adjective-Noun compositionality in English, languages exhibit compositionality in many ways. An important next step is to extend our analyses to other compositional spaces, such as Subject-Verb and Verb-Object combinations, to composition beyond two-word phrases, as well as to languages other than English.

Finally, our work provides a step towards evaluating the role of compositionality—long considered a central feature of language—and opens up future questions about what situations might make compositional language useful and crucial. We also highlight a challenging testbed for LLMs as models of compositional meaning, with avenues for building better models of human language processing that can account for unusual semantics.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback on our submission.

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Biemann, C., & Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality* (pp. 21–28).

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4427–4442).

Christiansen, M., & Chater, N. (2015). The language faculty that wasn't: a usage-based account of natural language recursion. , *6*.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159–190.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, *71*, 273–303.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *BioRxiv*, 2022–10.

Johnson, M. (2020). Compositionality. In D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, & T. Zimmermann (Eds.), *The wiley blackwell companion to semantics* (1st ed., pp. 1–27). Wiley. doi: 10.1002/9781118788516.sem094

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., … Lenci, A. (2023). Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), e13386.

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. In *Ninth conference of the european chapter of the association for computational linguistics* (pp. 30–36).

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., … others (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, E., Cui, C., Zheng, K., & Neubig, G. (2022). Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4437–4452).

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402.

Partee, B., et al. (1984). Compositionality. *Varieties of formal semantics*, *3*, 281–311.

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*, 447–471.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108).

Smith, K., & Kirby, S. (2012). Compositionality and linguistic evolution. In *Oxford handbook of compositionality.* Oxford University Press.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., … Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 1–18.

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. , *41*(1), 102–136. doi: 10.1111/cogs.12330