Language use is only sparsely compositional: The Case of English Adjective-Noun Phrases in Humans and LLMs

Aalok Sathe[1] **(asathe@mit.edu)**
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
McGovern Institute for Brain Research, Massachusetts Institute of Technology

Evelina Fedorenko[2] **(evelina9@mit.edu)**
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
McGovern Institute for Brain Research, Massachusetts Institute of Technology
Speech and Hearing Bioscience and Technology (SHBT) Program, Harvard University

Noga Zaslavsky[1,2] **(nogaz@uci.edu)**
Department of Language Science, University of California, Irvine

[1]corresponding author
[2]denotes co-senior authors

Abstract:

Compositionality is a 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

Introduction

Compositionality—the ability to combine units in language to produce novel meanings—is seen as a core principle that allowed human communication systems to flourish (Johnson, 2020; Smith & Kirby, 2012; Chaabouni et al., 2020). However, the extent to which compositionality prevails in human languages remains unclear: most research focuses on *item-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. In this work, we focus on the latter and estimate the utilization of the compositional capacity of language with English adjective-noun (Adj-N) pairs 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 (Study 1). Our results suggest that English vastly underutilized its compositional potential in this space. To control for finite-sample effects, we collect 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).

Study 1: System-level compositionality

We use information theory to quantify compositional capacity and its utilization. We take the joint entropy of adjectives and nouns, *H(A,N)*, as a measure of utilized compositionality. Using a well-known identity that relates entropy with mutual information (Cover, 1999): $H(A,N) = H(A) + H(N) - I(A;N)$, we see that if the marginal distributions are known and fixed, then the 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. To obtain these distributions, 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) as a summary of language use across 1991-2012. We observe 4.4M pairs with at least one occurrence. Given the finite nature of the corpus, it is unreasonable to expect to see all possible Adj-N combinations realized. Instead, we set an expectation based on the marginal distributions of adjectives and nouns treating them as independent random variables. We sample as many token pairs as are observed in the corpus: 25M. 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. MaxComb provides us with 8.9M (SD=1.75*k*, 100 repetitions), more than double the unique 4.4M observed in the corpus. Second, we use large language models (**LLMs**) as models of the compositional semantic space. LLMs are trained on much larger datasets than COCA: their data comes from varied sources including the Internet, and as such, they may offer a different account of typical language use than a corpus. To make our analyses and experiments tractable, we use only the 1,000 most frequent adjectives and nouns as observed in COCA, and the 1M theoretical combinations that can result. We observe a total of 11.7M combinations (0.3M unique) in this subspace in COCA. Next, we estimate LLM probabilities over sequences using the setup *"P_{LLM} (N | How likely is this: A)"*. We use several autoregressive models of varying numbers of parameters. Both COCA and MPT-30B—an LLM with 30B parameters—estimate a lot more sparsity in the

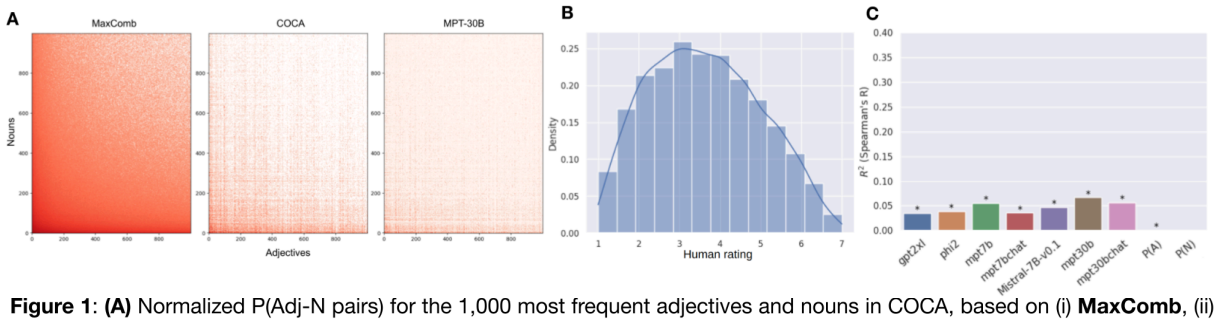

Figure 1: **(A)** Normalized P(Adj-N pairs) for the 1,000 most frequent adjectives and nouns in COCA, based on (i) **MaxComb**, (ii) **COCA**, and (iii) **MPT-30B** (adjusted to the COCA marginal adjective distribution). Both COCA and MPT-30B reflect substantially lower degree of compositionality compared to MaxComb, which affords the maximal degree of compositionality. **(B)** Human acceptability ratings for unattested Adj-N pairs, on a Likert scale from 1 (doesn't make any sense) to 7 (makes perfect sense). **(C)** LLMs align poorly with people's sensibility judgments on unattested stimuli. The y-axis denotes Spearman's correlation with people's average rating per item and $P_{LLM}(N|A)$. P(A), P(N) denote marginal COCA distributions. (*denotes $p < 0.001$).

compositional space than MaxComb suggests (Fig. 1A). We observe $I_{MPT\text{-}30B} > I_{COCA} > I_{MaxComb}$, with the mutual information of the baseline being 0 as expected (pairs are combined non-systematically).

Study 2: Are unattested Adj-N combinations sensible?

To disentangle finite-sample effects from true underutilization of compositional capacity we turn to the space of rare items. Seeing a high number of meaningful items in this space would suggest the corpus underestimates the utilization of compositionality. We collected sensibility judgments for 10,000 items—a subset of stimuli used by Vecchi et al. (2017). Unlike the original study, which used a forced-choice approach, we use Likert scale ratings as a way to get more direct per-item ratings with a relatively large number of raters per item. We recruited 1,000 fluent English-speakers on Prolific. Participants each rated 200 pairs based on "whether it makes sense on a scale from 1 (doesn't make any sense) to 7 (makes perfect sense)". In addition, we included 30 control items shared across all participants to compute inter-rater agreement. The mean split-half correlation with 1,000 bootstrapped iterations was 0.99. The mean leave-one-participant-out correlation was 0.82. Most pairs received low ratings, suggesting that the compositional capacity of Adj-N pairs is indeed underutilized. A large number of pairs also receive high ratings with high agreement (Fig. 1B), suggesting both the corpus and LLMs underestimate the utilization of compositional capacity.

Study 3: Do LLMs agree with humans in their judgments of unattested pairs?

Both LLMs and the corpus indicate a high amount of sparsity, and human judgments confirm that much of the compositional space is underutilized. However, high consistency across human raters suggests that humans can reason about the meanings of unattested compositional items. Given the recent success of LLMs and their ability to generalize beyond natural language statistics, we asked to what extent LLMs are good models of human judgments of unattested Adj-N combinations, thus focusing on linguistic input that is likely not represented in their training data. We find that LLMs are poorly aligned with human judgments on this set of items largely unattested in COCA. Neither LLMs nor the lexical frequencies of individual adjectives and nouns from the Adj-N pairs exceed an R^2 (Spearman's correlation) of 0.1 when compared with human judgments (Fig. 1C). The fact that human judgments cannot be predicted from the marginal lexical distributions alone points to the necessity of compositional semantic understanding in evaluating these Adj-N pairs. Poor alignment between LLMs and human judgments suggests that LLMs may not be good models of the compositional semantics that humans may be employing towards these items.

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 vastly underutilized. Human judgments confirm this observation by sampling the space of rarely-used Adj-N pairs. However, when it comes to these rare items, humans are highly consistent in their judgments, yet LLMs fail to capture them, suggesting a gap in the compositional semantics that can be induced solely from distributional data: formal linguistic competence (Mahowald, Ivanova et al., 2024) does not necessitate rich semantic understanding.

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.

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

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.

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

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