

## 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.