**Rich data drive generalization: lessons from machine learning for linguistics & cognitive science**

A commentary on Futrell & Mahowald's "How Linguistics Learned to Stop Worrying and Love the Language Models"

Andrew Kyle Lampinen
Google DeepMind

**Abstract: The diversity of variation captured in data can strongly affect the generalization of a learning system—even when that variation occurs along axes orthogonal to the generalization in question. Thus, I argue that data richness both distinguishes current language models from prior linguistic models, and may still underlie their remaining linguistic data inefficiency.**

Language model training data are distinct not only in their quantity. In comparison to classic linguistic and cognitive models, language model training data are fundamentally richer, more diverse, and much more noisy. Yet, they are still missing much of the diversity of human learning experiences. While Futrell & Mahowald discuss the role of data *quantity* in model training, they do not focus on the important role of data *richness*: that the diversity of *variation* captured in data can strongly affect the generalization of a learning system (cf. Raviv, Lupyan, & Green, 2022). This principle that learning diversity drives generalization has been frequently observed in machine learning, within and beyond the language domain (e.g., Yu et al., 2022; Goyal et al., 2022; del Rio et al., 2025).

One illustrative example of this can be found in some of the target authors' own work, though it is not emphasized as such in the target article: Misra & Mahowald (2024) show that under the same train-test *syntactic* generalization split, variability in the *semantic fillers* observed in training data affects the degree of *syntactic* generalization. That is, variation along non-syntactic axes can enhance generalization to novel syntactic structures. (Of course, seeing greater variability in syntactic structures will also typically improve generalization; e.g. Ahuja et al., 2025; Qin et al., 2025.)

Our prior work (Hill et al., 2020) on compositional generalization of language-learning agents shows a more dramatic effect: agents trained to perform classification through interaction in 3D environments show strong *compositional* language generalization, while agents trained to perform the same tasks on 2D images from the same environments generalize much more poorly. Thus, when comparing two language-learning systems — both of which are grounded, in the sense that their training tasks involve producing responses that depend on linking language to visual inputs and action outputs — the richness of the environmental grounding can significantly alter *linguistic* generalization. At a more basic level, various works have found that providing some type of semantic grounding can enhance syntactic generalization compared to systems that lack such grounding (e.g., Yedetore & Kim, 2024).

These works showing the contribution of data diversity to generalization in controlled experimental settings complement the many works showing its importance in practical language model training. For example, targeting sampling of training data for diversity can improve out-of-distribution perplexity (Yu et al., 2022). Conversely, language model training data collected from the internet often lack diversity; for example, Lee et al. (2022) highlight a single sentence that appears over 60,000 times in a standard training corpus. What might naively might seem to be 3.5 million tokens of unique training data (almost a year of a child's experience) turns out to be only 61 words of unique content — an important caveat to remember when seeing naive numeric comparisons of learning efficiency. While this is a particularly egregious example, the problem of partially-duplicated training data is prevalent, and Lee et al. (2022) show that reducing this duplication can substantially improve performance as well as training efficiency. This work provides another example of how learning diversity is important in practice as well as in theory: repeating data can actually harm performance.

These works illustrate how — both in controlled settings, and in practice when training language models — richer variation in learning experiences can fundamentally change the resulting patterns of generalization. Importantly, this is true both for variation along the axes on which generalization will be assessed (e.g. the diversity of syntactic structures seen), but also along axes of variation that seem orthogonal to the syntactic or compositional language generalization in question (e.g. the visual richness of multimodal inputs).

These findings have important theoretical implications for the linguistic and cognitive issues at play in the target article. In particular, they suggest that at least some of the benefit of data *quantity* for model performance may really be a benefit of data *diversity*, which only happens to be achieved through scaling the datasets. In itself, I believe this lesson is important for linguistics and cognitive science: how many famous negative results about neural networks (or other models) failing to generalize were wild extrapolations from training models on toy data that lacked all the richness of human experiences? We need better theories of how diversity contributes to generalization, and we should discount models that neglect the diversity of human learning experiences without a compelling reason.

Yet if language models benefit from data diversity, why do they still need so much more data than humans? Of course, many factors may contribute; choices like data-deduplication strategies, tokenization, and optimizers can all drastically change learning efficiency; we probably have not saturated improvements along any of these dimensions. However, I suggest that differences in the richness of the input humans experience may play a crucial role. Human linguistic inputs begin as auditory signals rather than discrete tokens; these signals carry much more information content to predict. Furthermore, human multimodal grounding is much richer. Indeed, while many models are now trained in part with multimodal inputs, for various reasons language still tends to dominate (e.g., Sim et al., 2025) — whereas for humans, language develops more slowly than other systems for interaction. Indeed, there have been many arguments that features like prosody (e.g., Gervain et al., 2020), gesture (e.g., Iverson & Goldin-Meadow, 2005), and joint attention (Tomasello & Farrar, 1986) play an important role in

human language development. Current language model training paradigms lack these, and many other rich features of human learning. However, simulating richer learning experiences is increasingly achievable. Thus, I believe that the modern paradigm of data-driven machine learning opens up many promising directions for exploring the role of linguistic and nonlinguistic data features in language acquisition (cf., Carvalho & Lampinen, 2025).

Acknowledgments:
I thank Felix Hill for his mentorship and many conversations on these issues, Michael Mozer for comments and suggestions, and Michael Terry for support.

References:

Ahuja, K., Balachandran, V., Panwar, M., He, T., Smith, N. A., Goyal, N., & Tsvetkov, Y. (2025). Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, 13, 121–141.

Carvalho, W., & Lampinen, A. (2025). Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.

del Rio, F., Raymond-Sáez, A., Florea, D., Icarte, R. T., Hurtado, J., Calderon, C. B., & Soto, A. (2025). Data Distributional Properties As Inductive Bias for Systematic Generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 25590-25601).

Gervain, J., Christophe, A., & Mazuka, R. (2020). Prosodic bootstrapping. *The Oxford Handbook of Language Prosody*, 563–573.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., ... & Bojanowski, P. (2022). Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020) Environmental drivers of systematicity and generalization in a situated agent. In *Proceedings of the International Conference on Learning Representations*.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science,* 16(5), 367–371.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022, May). Deduplicating training data makes language models better. In Proceedings of the 60th

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8424-8445).

Misra, K., & Mahowald, K. (2024). Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 913-929). Association for Computational Linguistics.

Qin, T., Saphra, N., & Alvarez-Melis, D. (2025). Sometimes I am a tree: Data drives unstable hierarchical generalization. *Advances in Neural Information Processing Systems*.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences,* 26(6), 462-483.

Sim, M. Y., Zhang, W. E., Dai, X., & Fang, B. (2025, July). Can vlms actually see and read? A survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 24452-24470).

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. Child development, 1454-1463.

Yedetore, A., & Kim, N. (2024). Semantic training signals promote hierarchical syntactic generalization in transformers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4059-4073).

Yu, Y., Khadivi, S., & Xu, J. (2022, October). Can data diversity enhance learning generalization?. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4933-4945).