Beyond the data gap: Children create languages, violate their input statistics, and exhibit critical periods

Commentary on Futrell & Mahowald: How Linguistics Learned to Stop Worrying and Love the Language Models

Annika McDermott-Hinman[1] & Roman Feiman[1,2]

1. Department of Cognitive and Psychological Sciences, Brown University
   Box 1821; Metcalf Research Building
   190 Thayer Street
   Providence, RI 02912
   +1 (401) 863-2727
2. Program in Linguistics, Brown University, Providence, RI 02912

Contact
Annika McDermott-Hinman: annikamh406@gmail.com; https://annikamh406.github.io
Roman Feiman: roman_feiman@brown.edu; https://sites.brown.edu/bltlab

Abstract word count: 60
Main text word count: 992
References word count: 736
Entire text word count: 1925

Abstract
Futrell & Mahowald argue LM success suggests humans may learn language entirely through domain-general statistical mechanisms. However, children differ crucially from LMs in their ability to surpass their input, their language learning trajectory, and the presence of a critical period. Until LMs account for these phenomena, it remains possible that human language acquisition is supported by innate, language-specific learning mechanisms.

Futrell and Mahowald argue that because language models (LMs) can learn linguistic structure from text input without explicitly encoded, language-specific inductive biases, we need not suppose children have such biases either. This argument requires that LMs learn linguistic structure in the same way as children; the only difference Futrell and Mahowald consider is how much data each needs to reach mature competence (see also Frank, 2023). Yet LMs and children also systematically diverge in *how* they acquire language, and even in what it is they acquire. If LMs cannot eventually model these aspects of acquisition, language-specific learning mechanisms may still be needed to explain them.

Language emergence. When children receive sparse linguistic input, they spontaneously generate linguistic structure not present in that input. Deaf children not exposed to sign language develop homesign systems whose structure and expressiveness outstrip the gestures of their hearing families. Homesign systems morphologically or syntactically distinguish parts of speech (Abner et al., 2019; Goldin-Meadow et al., 1994) and thematic roles (Goldin-Meadow & Mylander, 1998), and structurally mark grammatical subjects (Coppola & Newport, 2005)—distinctions absent or much reduced in their hearing families' gestures.

Further, when children's input is limited to a rudimentary linguistic system like homesign, they quickly expand and regularize it. Famously, the conventionalized sign system generated de novo by children attending a Nicaraguan school for the deaf was elaborated by successive student cohorts, each imposing additional structure absent from their input. For instance, the second cohort, but not the first, uses spatial modulation to convey shared reference (Senghas, 2003; Senghas & Coppola, 2001). Over a few cohorts, Nicaraguan Sign Language has come to exhibit discrete, compositional morphology (Senghas et al., 2004) and recursion (Kocab et al., 2023). Much of the structure of emerging languages originates from the child, not their input.

In contrast, LMs learn only to recapitulate the statistical properties of their training data. To account for language emergence—that is, the addition of linguistic structure that is absent from the input—LMs would need to *violate* those statistical properties. If that is possible, LMs would then need to more specifically enrich the structure of a language in the ways children do when placed under similar communicative pressures. Although simple neural agents can create rudimentary compositional systems in cooperative reference games (Lazaridou & Baroni, 2020; Steinert-Threlkeld, 2020; Boldt & Mortensen, 2024), it is uncertain whether and under which conditions LMs could create systems with the rich structural complexity that emerges in the ad-hoc communication system of even a single child.

Learning trajectories. Children's word-learning trajectories exhibit characteristic patterns distinct from the frequency of words in their input. Concrete nouns are over-represented at first, then verbs and adjectives, and only later do closed-class function words like "of" and "the" follow—though these are the most frequent words in the input (Frank et al., 2016). Further, even the most basic syntactic productivity (e.g., combining two words) emerges only after a period when children communicate meaningfully using early words in isolation (Fenson et al., 2007). LMs show the opposite pattern: they produce correct complex multi-word syntax early in training, but semantically coherent productions only much later (Müller-Eberstein et al., 2023; Xia et al., 2023). And unlike children, LMs' early productions reflect token input frequencies. "The", periods, and commas dominate (Chang et al., 2024).

These learning trajectories reflect fundamental differences in how children and LMs use language. LMs learn patterns of co-occurrence between constituents of their text-based input. Children must additionally learn how words *refer* to the world. This is why children first learn nouns for concrete objects: those are the easiest to identify as referents in their environment (Gillette et al., 1999; Snedeker et al., 2007). This gap might be bridged by multi-modal LMs that learn both word-word and word-percept relations. Yet early attempts show that such models have severe limitations in even the simplest referential abilities, such as generalizing a learned label-object pairing to new objects of the same shape (Vong et al., 2024). While more training data may help, these limitations may reflect deeper differences in how LMs and children acquire the concepts underlying word meanings. Just as evidence from language emergence shows that children add morphosyntactic structure absent from their linguistic input, scholars from Kant to Spelke to Gleitman have adduced evidence that children's preverbal conceptual and perceptual capacities structure thought beyond the combined associations of words with other words and with images. To the extent that learning language is the process of learning the expression of thought, LMs would need to model humans' capacities for thought to fully model their language.

Critical periods. While general statistical learning abilities improve throughout development (Shufaniya & Arnon, 2018), children's ability to learn the grammar of their language declines around puberty (Hartshorne, et al., 2018; Johnson & Newport, 1989; Lenneberg, 1967), and the ability to learn native phonology declines in infancy (see Werker, 2024, for review). This dissociation implies that learning language does not rely solely on general statistical mechanisms.

LMs might account for critical periods by demonstrating that declines in language learning ability can arise from language-specific experience that would not impede general statistical learning. However, recent work has found the opposite: learning a first language *accelerates* the LM's rate of subsequently learning a second (Oba et al., 2023). In fact, the way that a critical period *can* be induced in LMs is by imposing a regularizer partway through training—a process more analogous to a maturational decrease in language-specific plasticity than to any effect of experience (Constantinescu et al., 2025). By contrast, neural networks in other domains do show spontaneous critical periods. For instance, deep multi-sensor perceptual networks show a critical period during which improperly correlated data can permanently impair the model's learning (Kleinman et al., 2023). Thus, while neural networks may be able to account for critical periods in some areas of development based on experience alone, the way LMs have been able to account for them in language specifically favors a maturational, language-specific explanation.

Acknowledgements

Our thanks to Ann Senghas for insightful comments and suggestions on an earlier draft and to Marie Coppola for helpful discussion.

Competing interests

None.

Funding statement

This work was supported in part by NSF CAREER award #2442374 to R.F.

References

Abner, N., Flaherty, M., Stangl, K., Coppola, M., Brentari, D., & Goldin-Meadow, S. (2019). The noun-verb distinction in established and emergent sign systems. *Language*, *95*(2), 230–267. https://doi.org/10.1353/lan.2019.0030

Boldt, B., & Mortensen, D. (2024). *A Review of the Applications of Deep Learning-Based Emergent Communication* (No. arXiv:2407.03302). arXiv. https://doi.org/10.48550/arXiv.2407.03302

Chang, T. A., Tu, Z., & Bergen, B. K. (2024). Characterizing Learning Curves During Language Model Pre-Training: Learning, Forgetting, and Stability. *Transactions of the Association for Computational Linguistics*, *12*, 1346–1362. https://doi.org/10.1162/tacl_a_00708

Constantinescu, I., Pimentel, T., Cotterell, R., & Warstadt, A. (2025). Investigating Critical Period Effects in Language Acquisition through Neural Language Models. *Transactions of the Association for Computational Linguistics*, *13*, 96–120. https://doi.org/10.1162/tacl_a_00725

Coppola, M., & Newport, E. L. (2005). Grammatical Subjects in home sign: Abstract linguistic structure in adult primary gesture systems without linguistic input. *Proceedings of the National Academy of Sciences*, *102*(52), 19249–19253. https://doi.org/10.1073/pnas.0509306102

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual* (2nd ed.). Brookes Publishing Co.

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.08.007

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Goldin-Meadow, S., Butcher, C., Mylander, C., & Dodge, M. (1994). Nouns and Verbs in a Self-Styled Gesture System: What's in a Name? *Cognitive Psychology*, *27*, 259–319.

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*(6664), 279–281. https://doi.org/10.1038/34646

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1), 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Kleinman, M., Achille, A., & Soatto, S. (2023). Critical Learning Periods for Multisensory Integration in Deep Networks. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24296–24305. https://doi.org/10.1109/CVPR52729.2023.02327

Kocab, A., Senghas, A., Coppola, M., & Snedeker, J. (2023). Potentially recursive structures emerge quickly when a new language community forms. *Cognition*, *232*, 105261. https://doi.org/10.1016/j.cognition.2022.105261

Lazaridou, A., & Baroni, M. (2020). *Emergent Multi-Agent Communication in the Deep Learning Era* (No. arXiv:2006.02419). arXiv. https://doi.org/10.48550/arXiv.2006.02419

Lenneberg, E. H. (1967). *Biological foundations of language*. John Wiley & Sons.

Müller-Eberstein, M., Van Der Goot, R., Plank, B., & Titov, I. (2023). Subspace Chronicles: How Linguistic Information Emerges, Shifts and Interacts during Language Model Training. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13190–13208. https://doi.org/10.18653/v1/2023.findings-emnlp.879

Oba, M., Kuribayashi, T., Ouchi, H., & Watanabe, T. (2023). *Second Language Acquisition of Neural Language Models* (No. arXiv:2306.02920). arXiv. https://doi.org/10.48550/arXiv.2306.02920

Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. *Cognitive Development*, *18*(4), 511–531. https://doi.org/10.1016/j.cogdev.2003.09.006

Senghas, A., & Coppola, M. (2001). Children Creating Language: How Nicaraguan Sign Language Acquired a Spatial Grammar. *Psychological Science*, *12*(4), 323–328. https://doi.org/10.1111/1467-9280.00359

Senghas, A., Kita, S., & Özyürek, A. (2004). Children Creating Core Properties of Language: Evidence from an Emerging Sign Language in Nicaragua. *Science*, *305*(5691), 1779–1782. https://doi.org/10.1126/science.1100199

Shufaniya, A., & Arnon, I. (2018). Statistical Learning Is Not Age-Invariant During Childhood: Performance Improves With Age Across Modality. *Cognitive Science*, *42*(8), 3100–3115. https://doi.org/10.1111/cogs.12692

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting Over: International Adoption as a Natural Experiment in Language Development. *Psychological Science*, *18*(1), 79–87. https://doi.org/10.1111/j.1467-9280.2007.01852.x

Steinert-Threlkeld, S. (2020). Toward the Emergence of Nontrivial Compositionality. *Philosophy of Science*, *87*(5), 897–909. https://doi.org/10.1086/710628

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Werker, J. F. (2024). Phonetic perceptual reorganization across the first year of life: Looking back. *Infant Behavior and Development*, *75*, 101935. https://doi.org/10.1016/j.infbeh.2024.101935

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., & Stoyanov, V. (2023). Training Trajectories of Language Models Across Scales. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13711–13738. https://doi.org/10.18653/v1/2023.acl-long.767