COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13256

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13256

This article is part of the "Progress & Puzzles of Cognitive Science" letter series.

Large Language Models Demonstrate the Potential of Statistical Learning in Language

Pablo Contreras Kallens,[a] Ross Deans Kristensen-McLachlan,[b,c,d] Morten H. Christiansen[a,c,d,e]

[a]*Department of Psychology, Cornell University*
[b]*Center for Humanities Computing, Aarhus University*
[c]*Interacting Minds Centre, Aarhus University*
[d]*School of Communication and Culture, Aarhus University*
[e]*Haskins Laboratories*

Received 31 October 2022; received in revised form 14 January 2023; accepted 19 January 2023

Abstract

To what degree can language be acquired from linguistic input alone? This question has vexed scholars for millennia and is still a major focus of debate in the cognitive science of language. The complexity of human language has hampered progress because studies of language–especially those involving computational modeling–have only been able to deal with small fragments of our linguistic skills. We suggest that the most recent generation of Large Language Models (LLMs) might finally provide the computational tools to determine empirically how much of the human language ability can be acquired from linguistic experience. LLMs are sophisticated deep learning architectures trained on vast amounts of natural language data, enabling them to perform an impressive range of linguistic tasks. We argue that, despite their clear semantic and pragmatic limitations, LLMs have already demonstrated that human-like grammatical language can be acquired without the need for a built-in grammar. Thus, while there is still much to learn about how humans acquire and use language, LLMs provide full-fledged computational models for cognitive scientists to empirically evaluate just how far statistical learning might take us in explaining the full complexity of human language.

Keywords: Large language models; Artificial intelligence; Language acquisition; Statistical learning; Grammar; Innateness; Linguistic experience

Correspondence should be sent to Morten H. Christiansen, Department of Psychology, 228 Uris Hall, Cornell University, Ithaca, NY 14853, USA. E-mail: christiansen@cornell.edu

Whether learning language can be achieved from experience alone or whether innate knowledge of grammar is required has been a central problem in the study of language at least since Chomsky's (1959) seminal critique of Skinner (1957). In recent decades, the sides of this debate have broadly coalesced into two camps (see Dąbrowska, 2015, for a review). On the one hand, some approaches emphasize the role of domain-specific, rule-like representations or computations that are at least partially hard-wired and then tuned to an individual's linguistic environment during development (Chomsky, 1995; Jackendoff, 2011; Pinker, 1994; for recent examples, see Chomsky, 2017; Jackendoff & Audring 2019; Yang, Crain, Berwick, Chomsky, & Bolhuis, 2017). On the other hand, the broad field of usage-based approaches (Tomasello, 2009) has denied the necessity of innate, language-specific knowledge and asserts that language can be learned from experience through domain-general mechanisms such as statistical learning (Christiansen & Chater, 2016), abstraction (Lieven, 2014), and generalization (Goldberg, 2019).

Much of the debate has centered around the hypothesized poverty of stimulus (PoS; Chomsky, 1980). In broad terms, the PoS argument asserts that the scrappy and haphazard nature of linguistic experience makes it impossible to accurately generalize it to new grammatical utterances to the extent that adults do; therefore, language-specific knowledge *must* underlie their productive grasp of grammar. During the first wave of neural network modeling (Rumelhart & McClelland, 1986), so-called "connectionist models" attempted to reproduce various linguistic phenomena as counterarguments to the PoS argument. However, although reasonably successful, these models were extremely narrow in scope and scale, as recognized by their own proponents (Elman, 2005; McClelland, Hill, Rudolph, Baldridge, & Schütze, 2020), limiting the thrust of their challenge to more traditional nativist approaches.

We suggest that recent advances in language modeling have finally fulfilled the promise of connectionist models as "empirical tests of learnability claims" (Elman et al., 1996, p. 385). Indeed, we argue that these models provide an existence proof that the ability to produce *grammatical* language can be learned from exposure alone without language-specific computations or representations. Particularly, we are referring to transformer-based large language models (LLMs) like GPT-3 (Brown et al., 2020), Gopher (Rae et al., 2021), OPT (Zhang et al., 2022), BLOOM (BigScience Workshop, 2022), among others. Compared to the connectionist models of old, the LLMs are truly large networks with hundreds of billions of weights and tens of layers—but just like their modest precursors, they are also trained to predict the next word in a sentence. Some of the key differences between LLMs and their older kin are the way they process their input (i.e., in parallel instead of sequentially) and the presence of an attention layer that can encode and weight the dependencies between its components (Vaswani et al., 2017). Both have allowed networks to become deeper and larger without the vanishing error gradient that afflicts their older recurrent counterparts. These improvements moreover make it possible to train them faster and in parallel on much larger corpora.

The key point that we want to make is that the output of these models is, almost without exception, *grammatical*. Even when examining experiments that are designed to exhibit some of the model's shortcomings, such as *SubSimulatorGPT3* (https://www.reddit.com/r/SubSimulatorGPT3/), a forum where all posts and comments are outputs of GPT3, one cannot help but notice that the content is grammatically correct. Despite the deluge of generated

long-form text that the launch of ChatGPT (https://openai.com/blog/chatgpt/) has brought about, one is hard-pressed to find examples of ungrammatical sentences in English. Further underscoring our point, the examples provided by their harshest critics to highlight the models' shortcomings are always grammatical, even when they are arguing against the plausibility of LLMs as models of human language (e.g., Marcus, 2022a; Marcus & Davis, 2020). Instead, it has been clear that their limitations reside in the semantic and discourse levels of language production.

It is easy to overhype LLMs, so we want to be clear about what we are *not* claiming. First, we are not claiming that these models *understand* language. Interacting with the world was not part of their training or architecture, so their utterances do not have meaning, at least not in the sense that the same words uttered by a human do. Moreover, they are not language *users* in the same way as humans are, as that requires a large additional set of cognitive abilities that they clearly lack, especially those related to social interaction (see Christiansen & Chater, 2022). Similarly, we are not claiming that they are intelligent, sentient, or agentive, and we acknowledge that their output can be racist, sexist, or express other harmful biases. In fact, we agree that LLMs are akin to powerful statistical engines adept at detecting and generalizing the probabilistic patterns found in the large volume of text they have been exposed to (Bender, Gebru, McMillan-Major, & Schmitchell, 2021). What we *are* suggesting, though, is that LLMs, like GPT-3, can produce human-level grammatical language without a built-in grammar and that this has important theoretical implications for cognitive science.

Still, it might be objected that LLMs are just mimicking language using statistical patterns (e.g., Marcus, 2022b) and that LLMs merely reuse bits of the language they have memorized and generalize by extrapolating over past chunks of input (e.g., Pinker, 2022). But this is exactly what usage-based theories suggest is a key aspect of language learning and use (e.g., Christiansen & Chater, 2016; Goldberg, 2019; Lieven, 2014; Tomasello, 2009). Indeed, there is mounting evidence that memorizing, abstracting, and generalizing multiword expressions is precisely how humans learn and use language (see Contreras Kallens & Christiansen, 2022, for a review). Thus, this objection might be turned on its head: If the overwhelmingly grammatical language produced by LLMs can be explained by statistical learning and generalization, what is the need for an innate grammar? In other words, LLMs can be viewed as a usage-based answer to the PoS argument—at least when it comes to the production of grammatical language.

A more empirical objection to our argument could point to the adequacy of LLMs as models of human learners based on their scale, computational principles, or training data. This, of course, is work yet to be done. Nevertheless, some recent studies have suggested that the analogy is not as far-fetched as intuition would suggest. For example, in a next-word prediction paradigm, Goldstein, Zada, and Buchnik (2022a) found a remarkable overlap between the predictions made by GPT-2 and human participants when listening to a podcast. Moreover, they suggest that GPT-2 embeddings contain information useful for encoding predictive brain activity in language processing. In an even more recent preprint, Goldstein et al. (2022b) found a correspondence between GPT-2's per-layer activation and the time course of language processing using EcoG (i.e., recordings from electrodes implanted in the brain). Importantly, this processing parallel does not seem to be a function of the unrealistic

number of tokens to which LLMs are exposed during training: Hosseini et al. (2022) found that GPT-2 embeddings can be used to predict the activation of the language network in an fMRI neuroimaging study even when trained on 100 million tokens or the equivalent of the first 10 years of a child's language exposure (Gilkerson et al., 2017). All this considering that the scale of GPT-2 is orders of magnitude smaller than the most recent wave of LLMs (Dettmers et al., 2022). Thus, there are reasons, albeit preliminary, to believe in the adequacy of LLMs as models of at least some dimensions of language acquisition and processing.

In contrast to the connectionist models of 20 years ago, contemporary LLMs provide actual working models of full-blown language skills that can be explored experimentally. Through careful analyses of their output, we can assess just how much can be learned from the statistical regularities of the linguistic environment (Futrell et al., 2019; Wilcox, Futrell, & Levy, 2022). Some of this work has already been done in the context of encoder-only masked language models, such as BERT and its related descendants (Ettinger, 2020; Pandia & Ettinger, 2021; Rogers, Kovaleva, & Rumshisky, 2020). Their failures, such as with semantic coherence or pragmatics (Arehalli, Dillon, & Linzen, 2022; Dou, Forbes, Koncel-Kedziorski, Smith, & Choi, 2022; McClelland et al., 2020), are also interesting and point to other central tenets of usage-based theories such as the role of environmental contexts, developmental histories, cognitive machinery, and functional pressures in human language learning and use (Christiansen & Chater, 2022). The recent availability of LLMs thus allows for new systematic explorations into what can and cannot be learned purely from regularities in the input, with many key facets still underexplored and ripe for picking.

In summary, we believe that, even considering their limitations and allowing for a reasonable amount of skepticism around their true capabilities, LLMs have the potential to inform future work in the cognitive science of language. They can be viewed as working models of the potential of pure statistical learning of grammar based on prediction, memorization, generalization, and abstraction. Given their status as models, exploring their limits can offer new insights into how humans learn and use language to communicate. However, considering that the learnability-based argument for innate knowledge of grammar rests on the literal impossibility of this approach achieving the level of performance that LLMs have, their mere existence is already a game changer.

Acknowledgments

This research was supported in part by a New Frontiers Grant from the College of Arts and Sciences at Cornell University awarded to MHC.

References

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. https://doi.org/10.48550/arxiv.2210.12187

Bender, E. M., Gebru, T., McMillan-Major, A., & Schmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT 2021*, Canada (pp. 610–623).

BigScience Workshop. (2022). BLOOM. Hugging Face. Available at: https://huggingface.co/bigscience/bloom

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chomsky, N. (1959). Review of *Verbal behavior* by B. F. Skinner. *Language*, *35*, 26–58. https://doi.org/10.2307/411334

Chomsky, N. (1980). *Rules and representations*. Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: The MIT Press.

Chomsky, N. (2017). The language capacity: Architecture and evolution. *Psychonomic Bulletin & Review*, *24*, 200–203. https://doi.org/10.3758/s13423-016-1078-6

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., & Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. https://arxiv.org/abs/2204.02311

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2022). *The language game: How improvisation created language and changed the world*. New York: Basic Books.

Contreras Kallens, P., & Christiansen, M. H. (2022). Models of language and multiword expressions. *Frontiers in Artificial Intelligence*, *5*, 781962. https://doi.org/10.3389/frai.2022.781962

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, *6*, 852. https://doi.org/10.3389/fpsyg.2015.00852

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8 (): 8-bit matrix multiplication for transformers at scale. arXiv: https://arxiv.org/abs/2208.07339

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., & Choi, Y. (2022). Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland (Vol. *1*, pp. 7250–7274).

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Elman, J. L. (2005). Connectionist models of cognitive development: Where next? *Trends in Cognitive Sciences*, *9*, 111–117.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://aclanthology.org/2020.tacl-1.3

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN (Vol. *1*, pp. 32–42).

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*, 248–265.

Goldberg, A. (2019). *Explain me this*. Princeton, NJ: Princeton University Press.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D. … Hasson, U. (2022a). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4

Goldstein, A., Ham, E., Nastase, S. A., Zada, Z., Grinstein-Dabush, A., Aubrey, B., Schain, M., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., Levy, O., & Hasson, U. (2022b). Correspondence between the layered structure of deep language models and

temporal structure of natural language processing in the human brain. *BioRxiv*. https://doi.org/10.1101/2022.07.11.499562

Hosseini, E. A., Schrimpf, M. A., 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*. https://doi.org/10.1101/2022.07.11.499562

Jackendoff, R. (2011). What is the human language faculty? Two views. *Language*, *87*, 586–624.

Jackendoff, R., & Audring, J. (2019). The Parallel Architecture. In A. Kertész, E. Moravcsik, & C. Rákosi (Eds.), *Current approaches to syntax: A comparative handbook* (pp. 215–240). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110540253-008

Lieven, E. (2014). First language development: A usage-based perspective on past and current research. *Journal of Child Language*, *41*, 48–63. https://doi.org/10.1017/S0305000914000282

Marcus, G. F. (2022a). Deep learning is hitting a wall. Nautilus. Available at: https://nautil.us/deep-learning-is-hitting-a-wall-238440/. Accessed October 26, 2022.

Marcus, G. F. (2022b). Noam Chomsky and GPT-3 [Blog Post]. The road to AI we can trust. Available at: https://garymarcus.substack.com/p/noam-chomsky-and-gpt-3. Accessed October 26, 2022.

Marcus, G. F., & Davis, E. (2020). August 22nd). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about [Blog Post]. *MIT Technology Review*. Available at: https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. Accessed October 26, 2022.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*, 25966–25974.

Pandia, L., & Ettinger, A. (2021). Sorting through the noise: Testing robustness of information processing in pre-trained language models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic (pp. 1583–1596). https://aclanthology.org/2021.emnlp-main.119

Pinker, S. (1994). *The language instinct: The new science of language and mind*. William Morrow and Company.

Pinker, S. (2022). Pinker's initial salvo. Shtetl-Optimized: The Blog of Scott Aaronson. Available at: https://scottaaronson.blog/?p=6524. Accessed October 26, 2022.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van den Driessche, G., Hendricks, L. A., Rauh, M., Huang, P. -S., … Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. arXiv https://doi.org/10.48550/arXiv.2112.11446

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Rumelhart, D. E., McClelland, J. L., & Research Group, P. D. P. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Skinner, B. F. (1957). *Verbal behavior*. Princeton, NJ: Prentice-Hall.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. L. Bavin (Ed.), *The Cambridge handbook of child language* (pp. 69–87). Cambridge, MA: Cambridge University Press.

Wilcox, E. G., Futrell, R., & Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*. https://doi.org/10.1162/ling_a_00491

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30*, Long Beach, CA.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, *81*, 103–119.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models. arXiv https://doi.org/10.48550/arXiv.2205.01068.