https://doi.org/10.1038/s42256-023-00703-8

Language models and linguistic theories beyond words

Check for updates

The development of large language models is mainly a feat of engineering and so far has been largely disconnected from the field of linguistics. Exploring links between the two directions is reopening longstanding debates in the study of language.

Frederick Jelinek, a renowned Czech-American researcher in natural language processing and speech recognition, famously said in 1985, "Every time I fire a linguist, the performance of the speech recognizer goes up"[1], suggesting that there may be no efficient way to include linguistic knowledge in such systems[2]. Does this sentiment also hold true for state-of-the-art large language models (LLMs), which seem to be mostly artefacts of computer science and engineering? Both LLMs and linguistics deal with human languages, but whether or how they can benefit each other is not clear.

To start discussing connections between the two fields, a distinction needs to be made between computational linguistics and other kinds of linguistics – theoretical, cognitive, developmental and so on. Computational linguistics traditionally uses computational models to address questions in linguistics and borders the field of natural language processing, which in turn builds models of language for practical applications such as machine translation. The Annual Meeting of the Association for Computational Linguistics (ACL), the largest conference in the field, has seen an increase of 44% in the number of submissions over the past year, from 3,378 in 2022 to 4,864 in 2023. These numbers are hardly surprising given the rise of natural language processing in the past few years and, more recently, of LLMs. There is also increasing interest from researchers in other disciplines who recognize the potential of computational models of language in their own work. An article in our May 2023 issue proposes drawing inspiration from computational linguistics and natural language processing for building protein language models[3]. Another recent article in *Nature* uses a classical computational linguistic approach for designing mRNA vaccines[4].

But other linguistic disciplines, such as cognitive and developmental linguistics, which focus on child language acquisition and human cognition, are becoming more visible as well. For instance, in the search for computational models inspired by infant-like learning, researchers are considering the kind of input that babies learn from[5]. An exciting step in this direction is the BabyLM challenge, which gives machine learning researchers the task of training language models from scratch on amounts of linguistic data similar to those available to a 13-year-old child: around 100 million words, rather than the estimated 300 billion words ingested by ChatGPT.

It is generally agreed that LLMs do not implement a particular linguistic theory. Noam Chomsky, the pioneer of modern linguistics, likened LLMs to a bulldozer, saying that they are a useful tool to have but "not a contribution to science." Other scientists, however, hold a diametrically opposite view: Steven Piantadosi, a professor of psychology and neuroscience at the University of California, Berkeley, recently stated that LLMs are "precise and formal accounts" of language learning, and that their success brings Chomsky's influential linguistic theory of universal grammar, which postulates the existence of innate biological constraints that enable humans to learn languages, to "a remarkable downfall"[6]. Although this specific debate recently attracted media attention, it is reminiscent of other ongoing discussions in linguistics and cognitive science. One of them, which we brought up in our April 2023 editorial[7], is a debate on whether LLMs are truly capable of understanding language or merely mimic it[8]. Another dispute is between those who consider statistical pattern discovery to be a useful tool in linguistics and language acquisition, and those who, like Chomsky, think this sort of empirical analysis of surface language forms is fruitless and the only viable approach is to look at the underlying syntactic structures. Although there are nuances to such debates, all of them share a disagreement about how useful – for science, humanity and linguistics – the state-of-the-art LLMs are, and whether their cost is justified.

The positions taken by each side in these debates are often extreme, but there have also been more balanced views on what linguistics and state-of-the-art computer models can offer each other. Connections between theoretical linguistics and deep learning were discussed several years ago in *Language*, wherein Tal Linzen, a professor of linguistics and data science at New York University, highlighted possible pathways for interaction between deep neural networks and research on language. He argued that linguists could benefit in various ways from the platform for constructing models of language acquisition and processing that neural networks provide[9]. This recommendation may apply equally well, if not even better, to the recent LLMs.

From the cognitive perspective, a balanced view on the relationship between LLMs and human cognition was outlined in a recent preprint article inspired by research in neuroscience[10]. Although LLMs excel at language, they are not models of thought – or, in linguistic terminology, they succeed at formal competence, being able to generate meaningful and coherent texts and replicate some complex human-like linguistic behaviours, but fail at functional competence, which has to do with world knowledge and pragmatics. The balance, therefore, may lie in using LLMs in the capacity they actually possess: as language tools that can, for example, assist us in writing texts, translating them into a different language, generating code in programming languages, etc.

LLMs currently have little to do with linguistics and human cognition, and there is a chance that in the future they will diverge even more[11]. However, the field of linguistics is clearly affected by the development of tools so powerful that their output can easily be confused with human-generated texts. LLMs are again reopening some of the debates in linguistics that have been ongoing for decades[12], and there is hope that they will be put to good use in future linguistic research efforts.

Published online: 21 July 2023

References

1. Moore, R. K. *ISCA* https://www.isca-speech.org/archive/pdfs/interspeech_2005/moore05_interspeech.pdf (2005).
2. Jelinek, F. *Lang. Resour. Eval.* **39**, 25–34 (2005).
3. Vu, M. H. et al. *Nat. Mach. Intell.* **5**, 485–496 (2023).
4. Zhang, H. et al. *Nature* https://www.nature.com/articles/s41586-023-06127-z (2023).
5. Zaadnoordijk, L., Besold, T. R. & Cusack, R. *Nat. Mach. Intell.* **4**, 510–520 (2022).
6. Piantadosi, S. *LingBuzz* https://lingbuzz.net/lingbuzz/007180 (2023).
7. *Nat. Mach. Intell.* **5**, 331–332 (2023).
8. Mitchell, M. & Krakauer, D. C. *Proc. Natl Acad. Sci.* **120**, e2215907120 (2023).
9. Linzen, T. *Language* **95**, e99–e108 (2019).
10. Mahowald, K. et al. Preprint at https://doi.org/10.48550/arXiv.2301.06627 (2023).
11. Pater, J. *Language* **95**, e41–e74 (2019).
12. Piattelli-Palmarini, M. (ed.) *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky* (Harvard Univ. Press, 1980).