Large Language Models: The best linguistic theory, a wrong linguistic theory, or no linguistic theory at all?

To appear in Zeitschrift für Sprachwissenschaft
Draft

Comments welcome!

Stefan Müller

October 24, 2024

Abstract

This paper discusses Ambridge & Blything's claim (2024) that Large Language Models are the best linguistic theory we currently have. It claims that LLMs are wrong linguistic theories and concludes that they are not a linguistic theory at all. It is pointed out that Chomsky's claims about innateness, about transformations as underlying mechanisms of the language faculty and about plausible representations of linguistic knowledge are known to be flawed by quite some time by now and that we would not have needed LLMs for this. Chomsky's theories are not refuted by LLMs in their current form, since LLMs are different in many aspects from human brains. However, the tremendous success of LLMs in terms of applications makes it more plausible to linguists and laymen that the innateness claims are wrong.

It is argued that the use of LLMs is probably limited when it comes to typological work and cross-linguistic generalizations. These require work in theoretical linguistics.

1 Are Large Language Models linguistic theories?

In a recent paper in a special issue of Theoretical Linguistics containing "Reflections on Theoretical Linguistics" on the occasion of the 50th anniversary of the journal Ambridge & Blything (2024) claim that "large language models are better than theoretical linguists at theoretical linguistics" (p. 33). The authors examine the output of an LLM with regard to the argument structures of verbs, and are impressed that the model predicts the same as Ambridge found out in experiments with students. The authors claim that LLMs are a theory of language. The best one we have right now:

> large language models (LLMs) are already the leading current theories of how speakers learn and represent these restrictions. Of course, they are not perfect theories […] but they're better theories than any others that have been proposed. Ambridge & Blything (2024: 34)

LLMs are very interesting and you can do a lot of impressive things with them.[1] But are they theories? Do they help in any way to get a better understanding of language?

The authors claim that large language models are theories of language acquisition and representation and that they are instantiations of Construction Grammar (Goldberg 2006) approaches:

> Large language models [...] constitute theories of language acquisition and representation; theories that instantiate exemplar-, input- and construction-based approaches, though only very loosely. (Ambridge & Blything 2024: 33)

The authors claim that models are a theory (see also Piantadosi 2024: 360):

> OK, so the model makes the right predictions but – we hear you ask – where is the theory? That's the point: the model is the theory. (Ambridge & Blything 2024: 39)

This shows some confusion in terminology. Model-theoretic approaches assume that there is a theory and that the theory can be used to build models (Richter 2021). A theory should use primitives that are appropriate for a certain domain and it should contain statements about these primitives. Large language models are neuronal nets that have been organized and trained in a certain way. Nodes of such nets can be examined and we can even find certain information in them (Manning et al. 2020, Zhang & Bowman 2018) but this information is not a theory.

Later in the paper and contradicting their earlier claim, the authors argue that the programs that generate the LLMs are a theory:

> "But", critics object, "we have no idea what it's doing" (e.g., Kodner et al. 2023; Milway 2023). Quite the opposite: Unlike for traditional linguistic theories, every last detail of the model's assumptions and operation is written out in black and white, in thousands of lines of computer code. This code is a theory of the acquisition of (among other things) verb argument structure; it's even – like traditional linguistic theories – written in a language, albeit an artificial programming language, rather than a natural language like English. We

[1] For example, ChatGPT can explain prime factorizations in Trump-style (Piantadosi 2024: 356–357).

> know exactly what the model is doing. (Ambridge & Blything 2024: 39–40)

This very quote is an instance of mixing levels. We know what the *code* is doing. We do not know what the trained net, the model, is doing. This depends on the training data and even if we knew the training data, we could not predict what specific nodes in the net would do. The issue is just too complex for us humans and the training data is too vast. This can be compared with Definite Clause Grammars: this is a notation that can be used to write down phrase structure grammars. Most Prolog interpreters come with a component that parses such grammars directly (see Clocksin & Mellish 1984: Chapter 9 for more information on DCGs and Müller 2023c: Task 10 on p. 81 for more information on working with DCGs online). Clocksin & Mellish (1984: 268–270) provide two pages of code for the translation of DCGs into Prolog code. The resulting Prolog program does Parsing as Deduction. In this case, we know what the code is doing. It reflects our theory about language. For a more elaborate example of Parsing as Deduction based on Government & Binding see Johnson (1989).

Fox & Katzir (2024) published a response to Ambridge & Blything (2024) in the same issue of Theoretical Linguistics. They write:

> The distinction between competence and performance and between correctness and likelihood are parts of all the best theories of human linguistic cognition, as are the aspects of linguistic representation that we briefly reviewed (modularity, constituency, and entailment). […] the LLM Theory does not even come close to approximating the relevant observations. Obviously it cannot derive these properties of human linguistic cognition and without doing so it cannot be considered a scientific theory at all. (Fox & Katzir 2024: 75)

The authors claim that LLMs cannot be a theory, since they do not make the competence-performance distinction, since they do not adhere to modularity and since they do not capture constituency. If these failures to capture certain properties of language would indeed entail that LLMs are not scientific theories, then neither Construction Grammar nor any flavor of Mainstream Generative Grammar (MGG)[2] would be. The distinction between competence and performance is rejected in Usage-Based Construction Grammar (Diessel 2015: 297). I personally think that this is a mistake (Müller 2023c: Chapter 15), but nevertheless approaches in Usage-based Construction Grammar are theories. The alternative approaches in MGG do not fare any better. All basic architectural assumptions in all of Chomsky's approaches are highly implausible from a psycholinguistic point of view. The Derivational Theory of Complexity, which assumed that sentences involving more transformations in their analysis are more difficult to process than sentences with fewer transformations has been proven wrong (Fodor et al. 1974: 320–328, Müller 2023c: Chapter 15.1). The T-model

[2]The term MGG goes back to Culicover & Jackendoff (2005: 3). It refers to all proposals developed by Chomsky. Government & Binding (Chomsky 1981) and theories developed under the label of Minimalism (Chomsky 1995) are the most recent incarnations.

with its autonomous components of syntax, phonological and logical form has been proven wrong resulting in spectacular analyses in Cartography (Cinque & Rizzi 2010) to circumvent the autonomy of syntax restriction (see Müller 2023c: Section 4.6.1.1, 2023b: Section 4.10.2 on this point and on problems with Cartographic approaches, for example, Cinque's (1994: 100) claim that categories like Nationality are part of our genetic endowment). Derivation by Phase (Chomsky 2008) and other Minimalist architectures (Richards 2015: 812, 830) are entirely implausible as architectures for human language (Borsley & Müller 2021: Section 3.6), since they are incompatible with incremental parallel processing of linguistic information at all descriptive levels (Marslen-Wilson 1975, Tanenhaus et al. 1996). If the argumentation by Fox & Katzir (2024) was valid, it would follow that all approaches in Usage-based Construction Grammar and MGG were not scientific theories. This would be a very strange conclusion, but it is not warranted. They are scientific theories, but they are bad ones.

Concerning the other points raised in the above quote: the claims about modularity and interfaces are probably wrong (Pulvermüller 1999, Pulvermüller et al. 2013, Jackendoff 2000: 22, 27, Kuhn 2007) and there are theories in which constituency does not play a role but dependency does (Tesnière 1959). And Clark et al. (2019), Hewitt & Manning (2019), Manning et al. (2020) show that dependency information is encoded in LLMs.[3]

So Fox & Katzir (2024) argue that the theory, if it existed in LLMs, would be wrong. I argue that there is no theory about language in it. I believe that Ambridge & Blything (2024) are fundamentally wrong. To show this let us do a thought experiment. LLMs are neural networks. Their architecture is inspired by what we find in brains. They differ from brains in various ways, but let us assume that one could develop a perfect replica of a brain one day. To quote Norbert Wiener, the founder of cybernetics: "The best model of a cat is a cat, preferably the same cat."[4] So, if we have a perfect copy of somebody's brain, what can we do with it? The artifical brain can then do exactly what the 48 Liverpool students mentioned in the Ambridge & Blything (2024) paper can do.

[3] It is important to note that Clark et al. (2019), Hewitt & Manning (2019), Manning et al. (2020) were able to discover the fact that dependencies are represented in LLMs because they knew the concept of dependencies, which was developed by Tesnière in 1924–1954. So the linguistic theory and related concepts were a prerequisite to find linguistic structure in the neural networks. This point will be taken up again below in the discussion of typological work.

Another note on linguistic information in LLMs: Imagine you build a model of a landscape in a lab. You have soil and water. The water runs in little rivers, carves valleys into the soil. The landscape is formed over time, you get hills, canyons, creeks, rivers. This is like the training phase of a neuronal net. After this landscape forming phase you may put liquid into your artificial landscape and see what forms rivers will take. But does this tell you something about rivers in general? A theory about the way water distributes? No. It gives you concrete examples of how a possible river may look like after years and years of forming an artificial landscape. This is what we get from LLMs: we train them with lots and lots of data and then get a structure that was shaped by the data.

[4] To be clear about this: This was a joke. If your homework is to create a model of the cat Molly and you hand in Molly, your supervisor will not be amused. You will fail when it comes to the point when your model has to be compared to the original Molly. Similarly, handing in Daisy, the cat of your neighbor, will be considered cheating as well.

Perhaps a bit more smoothed, because this replica can be fed much, much, much more data than all Liverpool students will ever see in their 48 lives combined. Now the question is: What does this mean for linguistics? Is a replica of a brain a theory about language? No. It is a masterpiece of engineering. Nothing more. To build such masterpieces, you need theories about how brains work. You can then take parts of these theories and use them to build artificial brains. The code that people write to train the data structures they have created is code that is motivated by theories about the brain. It is not a theory, and certainly not a theory about language. The criticism that Ambridge & Blything (2024) reject is justified: LLMs are not theories about language; the information contained in LLMs is accessible only indirectly. Just as you cannot directly access the information in brains. You can only research the behavior of people. That's what we've been doing for hundreds of years. We look at what people say and write. We conduct experiments with people. We ask them about the acceptability of sentences. We test where they look when certain sentences are uttered (Tanenhaus et al. 1996). We measure brain activity (event-related potentials, cerebral bloodflow, etc.) We investigate what happens when certain areas of the brain are damaged. This gives us information about the processes and representations of linguistic knowledge in the brain. From this we can then draw conclusions for plausible theories.

What is it like with LLMs? They are like brains: black boxes. We could start playing around with them now and try to find out what is stored where and how. But what good would that do? Actually such a research field exists already. Bender & Koller (2020: 5185) call it "BERTology":[5] Engineers and linguists are playing with LLMs and check what they can do. This is interesting, but irrelevant for linguistics.[6]

Conclusion: We (as humanity) have created a technical masterpiece, but we know no more about our cognitive abilities than we did before.

[5] BERT stands for *bidirectional encoder representations from transformers* and is a shorthand for a large language model introduced by Google.

[6] This was a bit of a hyperbole. LLMs may be used to play around with data and to check what these models need as input to get certain facts about language right. This can help linguists to discover relations and dependencies between linguistic phenomena that are plausible parts of a linguistic theory (generalizations, constructions, families of constructions and the relations between constructions). For example, Misra & Mahowald (2024) show that LLMs perform above chance on phrases like *a five beautiful days*, provided certain other constructions are in the training corpus. So, the place of LLMs in linguistics seems to be the one of subjects that one can feed arbitrary training material and that one can interrogate without them getting tired and without the need of an ethics vote. Since LLMs are different from real humans, the resulting theories should be checked with reference to actual data and actual human behavior, but they can serve as a first inspiration.

2 Large Language Models, human first language acquisition and "Chomsky's approach to language"

Maybe the last sentence needs a bit of qualification. Piantadosi (2024) claims that Chomsky's approach to language has failed, that it was proven wrong by Large Language Models. As Piantadosi (2024: 366) writes himself LLMs "are trained on truly titanic datasets compared to children, by a factor of at least a few thousand". So, if linguistics is dealing with human capabilities, we are not quite there yet. To model language acquisition, we would need grounded input, we would need a realistic amount of training data, we would have to simulate the development of brains and the growth of cognitive capacities in early childhood.[7] But what the success of LLMs suggests is that an elaborated component of Universal Grammar is not needed, that the argument of the Poverty of the Stimulus was flawed and so on. Above I wrote that "we know no more about our cognitive abilities than we did before". And this is true. We knew in 1974 (50 years ago) that transformations are psycholinguistically implausible (Fodor, Bever & Garrett 1974: 320–328). Psycholinguists sympathetic with the Chomskyan paradigm suggested that we have our linguistic knowledge represented as a Transformational Grammar, but that it then gets compiled out into a set of templates that are equivalent ot the constructions of Construction Grammar (Frazier & Clifton 1996: 27). But this of course begs the question why one should not work in Construction Grammar or a related framework like Constructional HPSG (Sag 1997, Müller et al. 2024) from the beginning. What is the evidence for some underlying transformation-based representation of linguistic knowledge? The various architectures that were proposed over the years were psycholinguistically implausible too. The T-Model (Chomsky 1981, 1986) was implausible (Müller 2023c: Section 15.2) and this got only worse with Phase-based variants of Minimalism (Chomsky 1995, 2008, Richards 2015: 812, 830, Borsley & Müller 2021: Section 5). But if the theories are incompatible with empirical facts like incremental processing, how can they tell us anything about human cognition and inateness? The Principle & Parameter model of language acquisition Chomsky (1981: 6) failed in various respects. It was assumed that one parameter was related to many properties of a language and worked like a switch (Chomsky 2000: 8), but none of the suggested correlations held up (Haider 2001: Section 2.2, Müller 2023c: Section 16.1). The way parameterization was conceptualized was biologically implausible. For example, it was assumed that Subjacency was a universal principle and the parameterization concerned the part of speech of certain bounding nodes within nonlocal dependencies (Chomsky 1986: 40, Baltin 1981). First, it could be shown that subjacency does not hold in Dutch, German and English (Koster 1978: 52, Müller 1999: 211, 2004, 2007: Section 3, Strunk & Snider 2013) and second,

[7] Children regularize more than adults (Hudson & Newport 1999, Hudson Kam & Newport 2005), a fact that can be traced back to their limited brain capacity ("less is more"-hypothesis, Newport 1990).

it is biologically absolutely implausible that part of speech information is encoded in our genes (Bishop 2002, Fisher & Marcus 2005: Section 6.4.2.2). This was realized by Hauser, Chomsky & Fitch (2002). What remained as property that was assumed to be part of Universal Grammar was Merge, an operation for combining linguistic material. Somehow a triviality (Müller 2023c: 475). A triviality that caused another linguistic war (Pullum 2024).

There is one important aspect of research in the Principles & Parameters era: The systematic search for universals, for commonalities and differences lead to a much improved knowledge about variation. We know much more about language as such, that is, about structures that are similar in principle. For example, the German sentence in (1) is parallel to the English translation.

(1)	dass	die	Straßenbahnen	um	die	Ecke	quietschen
	that	the	trams	around	the	corner	squeak

‘that the trams squeak around the corner’

As Müller (2013) pointed out, it is possible to develop analyses that capture the commonalities although the linearization of the constituents differs in German and English (English is an SVO language and German is SOV). Typological research is fascinating and requires the comparison of many very different languages on a theoretical level. I doubt that the results of cross-linguistic research can be derived from LLMs, without any interaction with theoretical linguistics. Training LLMs on multilingual material will be non-trivial[8] and discovering cross-linguistic generalizations in network representations is probably impossible without a theoretically informed clue on what to look for (see also footnote 3). A suggestion for a methodological clean way of deriving cross-linguistic generalizations that differs from the MGG approach is assumed in the CoreGram project (Müller 2015).

Chomsky claimed that there would be language universals but there are no plausible candidates for syntactic universals left (Evans & Levinson 2009, Müller 2023c: Section 13.1). There are tendencies, for sure, but this is not sufficient for positing innate knowledge of language. Recently, I discovered a universal, but it is not syntactic but rather on the text level: the Festschrift Universal. The reader is referred to Müller (2024) for details.

The strongest argument for innate linguistic knowledge seemed to be the Argument from the Poverty of the Stimulus, but it was never actually correctly carried out (Pullum & Scholz 2002, Scholz & Pullum 2002). Chomsky repeated his favourite argument with question formation as auxiliary inversion throughout several decades (Chomsky 1971: 29–33, 2013: 39). Bod (2009) showed how frequencies of subtrees can be used to learn structures of auxiliary inversion even though the examples that Chomsky (wrongly[9]) claimed to be

[8] See Chang et al. (2024) for comments on the low quality of multilingual language models. Note also that a lot of typologically interesting languages are low-resource languages, so a massive training like with LLMs is not possible because of the lack of data. See Chang et al. (2024) on monolingual models for 350 languages.

[9] See Pullum & Scholz (2002: 41–45).

non-existent in the data were not used in the learning procedure. Chomsky ignored these insights (Berwick, Pietroski, Yankama & Chomsky 2011, Chomsky 2013: 39) and so we find the auxiliary inversion claim again two years later in the same journal that also published Bod's paper. Similarly pattern-based modeling language acquisition research was much more successful in explaining cross-linguistic differences in acquisition than alternative accounts couched in Chomskyan frameworks (Freudenthal et al. 2007).

Connected to the assumption of Universal Grammar is the assumption of a Core/Periphery distinction (Chomsky 1981: 7–8). The idea is that there is a core of linguistic knowledge that is determined by our genetic endowment and there is a periphery (e. g. idioms) that is learned in another way. The interesting and very simple argument against this is this: If we can learn the idiosyncratic parts of a language that is assigned to the periphery, we should be able to learn the more regular parts of the assumed core (Abney 1996: 20, Goldberg 2006: 14, Newmeyer 2005: 100, Tomasello 2006: 20, Müller 2014). See Müller (2014) and the CoreGram project (Müller 2015) for a method for deriving language-internal and also cross-linguistic generalizations and the notion of *Kernigkeit* (coriness) that does not refer to Chomsky's core/periphery distinction.

So, we knew that Chomskyan approaches to language and language acquisition failed in terms of their basic assumptions (transformations), they failed in terms of their architecture with respect to psycholinguistic evidence (separation of syntax and phonology and semantics in various forms) and they failed in respect to assumptions about genetics. Everybody working in non-Chomskyan paradigms knows this for more than a decade (see the first editions of my *Grammatical theory* textbook from 2010 and 2014 for a summary of the respective discussions in German and English, respectively). We did not need LLMs for this, but maybe the actual usefulness of these networks is that they make the possibility that we do not need any innate domain-specific knowledge plausible to everybody: linguists and laymen. However, to show that LLMs can acquire languages like humans do, they have to be more human-like. To reach this goal, we probably need more knowledge about brains. As I pointed out above, if we manage to reach the goal of creating more human-like models, we know how brains work, but we do not necessarily know how languages work.

3 Linguistic theories

I believe that linguistic theories should contain rules and symbols. A linguistic theory can to some extent be derived from large corpora using automatic methods. Both the categories can be obtained via class formation and rules or valence patterns and the corresponding lexical entries can be derived automatically. The parts of speech and features like case, gender and number that are currently used in linguistic theories are basically the outcome of a distributional analysis that was done "by hand" during the last centuries. Grammar rules and also feature-value pairs may be assigned probabilities (Jurafsky 1996). These can also be derived from corpora. This is complicated and the mathematics is not fully understood yet. But one can train the system on large amounts of data. The training procedure contains assumptions about language: there are categories, there are constituents. There will be a residuum of infrequent phenomena that will not be captured this way (for example apparent multiple frontings, see Müller 2003). Some fine-tuning will be required and this is where the linguist comes in: rare data and complicated interacting phenomena may decide between various alternative theories of a language (Müller 2023a: Chapter 6).

What would be missing in such grammars is the meaning component: a distributional analysis provides one with distribution classes, with syntactic regularities of the language under consideration. This is true for LLMs and any other outcome of a distributional analysis unless semantic information is explicitly encoded in the input and linked to real world experiences.[10] Therefore it is really surprising to see Construction Grammarians praising large language models as theories of human language. Wasn't it Construction Grammariens who told everybody in the field that human cognition is grounded (Barsalou 2008) and that language is not just abstract syntax and cannot be learned as such without a connection to semantics and the real world (Klein 1986: 44, Tomasello 2003: 113, Ambridge & Lieven 2011: Section 4.2.3, 4.2.8)? With grasping the communicative intention and attention sharing? Klein pointed out in 1986 already that no human being could learn Chinese by sitting in a room continually exposed to Chinese from loudspeakers.[11] This just would not work. But this is how LLMs learn: they just see masses and masses of text. BERT was trained by guessing masked words in a sentence and by guessing the next sentence. Children do not play such games. In fact, they have to solve a very hard puzzle on their own: the segmentation of the speech signal. They have to find out what the units are in order to be able to discover what they mean. As Bender & Koller (2020) pointed out: BERT and ChatGPT and the like do not have a clue about what they are "saying". Their representations do not have any connection to semantics, they are not grounded (Beuls & Van Eecke 2024).

[10]There is semantic knowledge implicit in LLMs. Piantadosi (2024: 358) points out that it is interwoven with syntactic information. The important point when it comes to human cognition is that the semantic knowledge in LLMs is not grounded. Jones et al. (2024: 2) discuss sentence-picture verification tasks. For example, hearers can infer from the sentences "He hammered the nail into the wall." and "He hammered the nail into the floor." that the nail is horizontal in the situation described by the first sentence and vertical in the second. This information is not explicitly coded in the sentences, so LLMs, which are trained on language alone, cannot learn this unless it is made explicit elsewhere in the training material.

See Chang & Maia (2001) for computational experiments on language acquisition with grounding in the framework of Construction Grammar and Steels (2003) for experiments with grounded communication in robotics. Beuls & Van Eecke (2024) extensively discuss the shortcomings of LLMs that are due to there representations not being grounded and they suggest ways to model grounded language acquisition. Jones et al. (2024) discuss first experiments with Multimodal LLMs and point out some shortcomings of current architectures.

[11]Klein speculated that at most phonological regularities can be learned and Newport et al. (2004) showed that humans can detect regularities by just being exposed to a continuous speech stream of syllables of various forms. Søgaard (2023: 44) pointed out that two year old infants can learn from TV, but TV involves another modality, the language is grounded (Rice 1983).

ChatGPT is a bullshit machine in the sense of Hicks et al. (2024), it is not and it does not contain a linguistic theory, not even a wrong one.

4 Conclusion

Large Language Models are not theories of language. To build LLMs, one needs a theory and depending on the goal to be reached, the theory may be a theory of the human brain. Knowing how a brain is working does not entail knowledge about language. To do typological research means to compare thousands of languages. This is done by theoretical linguists on a meta level and not within neuronal nets trained with input of thousands of languages. Of course, one can imagine typological research supported by computers, but it would require trained linguists who know what to look for. The existence and success of LLMs does not entail that the problem of human language acquisition is solved, since the architecture and the training process of LLMs is quite different from how human brains develop and how humans acquire language. However, LLMs show that the data is rich and make it even more plausible that humans are not born with innate domain specific knowledge about language.

Acknowledgements

I thank Rui Chaves, Mark Felfe, Hubert Haider, Joachim Jacobs, Tibor Kiss, Alexander Koller, Roland Schäfer, Oliver Schallert and Remi van Trijp, Giuseppe Varaschin for comments and discussion and Konstantin Schulz, Anke Lüdeling, and Martin Klotz for discussion and Tine Mooshammer for raising questions about Ambridge & Blything (2024).

Oliver Schallert suggested adding more discussion of Piantadosi's paper. Thanks for this, I think it really improved the paper.

References

Abney, Steven P. 1996. Statistical methods and linguistics. In Judith L. Klavans & Philip Resnik (eds.), *The balancing act: Combining symbolic and statistical approaches to language*, 1–26. Cambridge, MA: MIT Press.

Ambridge, Ben & Liam Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics* 50(1–2). 33–48.

Ambridge, Ben & Elena V. M. Lieven. 2011. *Child language acquisition: Contrasting theoretical approaches*. Cambridge, UK: Cambridge University Press.

Baltin, Mark. 1981. Strict bounding. In Carl Lee Baker & John J. McCarthy (eds.), *The logical problem of language acquisition*, 257–295. Cambridge, MA: MIT Press.

Barsalou, Lawrence W. 2008. Grounded cognition. *Annual Review of Psychology* 59. 617–645.

Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics.

Berwick, Robert C., Paul Pietroski, Beracah Yankama & Noam Chomsky. 2011. Poverty of the Stimulus revisited. *Cognitive Science* 35(7). 1207–1242.

Beuls, Katrien & Paul Van Eecke. 2024. Humans Learn Language from Situated Communicative Interactions. What about Machines? *Computational Linguistics*. 1–35.

Bishop, Dorothy V. M. 2002. Putting language genes in perspective. *TRENDS in Genetics* 18(2). 57–59.

Bod, Rens. 2009. From exemplar to grammar: Integrating analogy and probability in language learning. *Cognitive Science* 33(5). 752–793.

Borsley, Robert D. & Stefan Müller. 2021. HPSG and Minimalism. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*, 1253–1329. Berlin: Language Science Press.

Chang, Nancy C. & Tiago V. Maia. 2001. Grounded learning of grammatical constructions. In *Papers from the 2001 AAAI Spring Symposium on Learning Grounded Representations*. AAAI.

Chang, Tyler A., Catherine Arnett, Zhuowen Tu & Benjamin K. Bergen. 2024. *Goldfish: monolingual language models for 350 languages*. arXiv:2408.10441v1.

Chomsky, Noam. 1971. *Problems of knowledge and freedom*. New York, NY: Pantheon Books.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris Publications.

Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2000. *New horizons in the study of language and mind*. Cambridge, UK: Cambridge University Press.

Chomsky, Noam. 2008. On phases. In Robert Freidin, Carlos P. Otero & Maria Luisa Zubizarreta (eds.), *Foundational issues in linguistic theory: Essays in honor of Jean-Roger Vergnaud*, 133–166. Cambridge, MA: MIT Press.

Chomsky, Noam. 2013. Problems of projection. *Lingua* 130. 33–49.

Cinque, Guglielmo. 1994. On the evidence for partial N movement in the Romance DP. In Guglielmo Cinque, Jan Koster, Jean-Yves Pollock, Luigi Rizzi & Raffaella Zanuttini (eds.), *Paths towards Universal Grammar: Studies in honor of Richard S. Kayne*, 85–110. Washington, D.C.: Georgetown University Press.

Cinque, Guglielmo & Luigi Rizzi. 2010. The cartography of syntactic structures. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 51–65. Oxford: Oxford University Press.

Clark, Kevin, Urvashi Khandelwal, Omer Levy & Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov & Dieuwke Hupkes (eds.), *Proceedings of the 2019 acl workshop blackboxnlp: analyzing and interpreting neural networks for nlp*, 276–286. Florence, Italy: Association for Computational Linguistics.

Clocksin, William F. & Christopher S. Mellish. 1984. *Programming in Prolog.* 5th edn. Berlin: Springer-Verlag.

Culicover, Peter W. & Ray Jackendoff. 2005. *Simpler Syntax.* Oxford: Oxford University Press.

Diessel, Holger. 2015. Usage-based Construction Grammar. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handbook of cognitive linguistics*, 296–322. Berlin: Mouton de Gruyter.

Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *The Behavioral and Brain Sciences* 32(5). 429–448.

Fisher, Simon E. & Gary F. Marcus. 2005. The eloquent ape: Genes, brains and the evolution of language. *Nature Reviews Genetics* 7(1). 9–20.

Fodor, Jerry A., Thomas G. Bever & Merrill F. Garrett. 1974. *The psychology of language: An introduction to psycholinguistics and Generative Grammar.* New York, NY: McGraw-Hill Book Co.

Fox, Danny & Roni Katzir. 2024. Large language models and theoretical linguistics. *Theoretical Linguistics* 50(1–2). 71–76.

Frazier, Lyn & Charles Clifton Jr. 1996. *Construal.* Cambridge, MA: MIT Press.

Freudenthal, Daniel, Julian M. Pine, Javier Aguado-Orea & Fernand Gobet. 2007. Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science* 31(2). 311–341.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Haider, Hubert. 2001. Parametrisierung in der Generativen Grammatik. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*, vol. 1, 283–293. Berlin: Mouton de Gruyter.

Hauser, Marc D., Noam Chomsky & W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598). 1569–1579.

Hewitt, John & Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran & Thamar Solorio (eds.), *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)*, 4129–4138. Minneapolis, MI: Association for Computational Linguistics.

Hicks, Michael Townsen, James Humphries & Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology* 26(2). 1–10.

Hudson, Carla L. & Elissa L. Newport. 1999. Creolization: Could adults really have done it all? In Annabel Greenhill, Heather Littlefield & Cheryl Tano (eds.), *Proceedings of the Boston University Conference on Language Development*, vol. 23, 265–276. Somerville, MA: Cascadilla Press.

Hudson Kam, Carla L. & Elissa L. Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1(2). 151–195.

Jackendoff, Ray. 2000. Fodorian modularity and representational modularity. In Yosef Grodzinsky, Lewis P. Shapiro & David Swinney (eds.), *Language and the brain: Representation and processing*, 3–30. San Diego: Academic Press.

Johnson, Mark. 1989. Parsing as deduction: The use of knowledge of language. *Journal of Psycholinguistic Research* 18(1). 105–128.

Jones, Cameron, Benjamin Bergen & Sean Trott. 2024. Do multimodal large language models and humans ground language similarly? *Computational Linguistics*.

Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2). 137–194.

Klein, Wolfgang. 1986. *Second language acquisition*. Cambridge, UK: Cambridge University Press.

Koster, Jan. 1978. *Locality principles in syntax*. Dordrecht: Foris Publications.

Kuhn, Jonas. 2007. Interfaces in constraint-based theories of grammar. In Gillian Ramchand & Charles Reiss (eds.), *The Oxford handbook of linguistic interfaces*, 613–650. Oxford: Oxford University Press.

Manning, Christopher D., Kevin Clark, John Hewitt & Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* 48(117). 30046–30054.

Marslen-Wilson, William D. 1975. Sentence perception as an interactive parallel process. *Science* 189(4198). 226–228.

Misra, Kanishka & Kyle Mahowald. 2024. *Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs*. `https://arxiv.org/abs/2403.19827`.

Müller, Stefan. 1999. *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*. Tübingen: Max Niemeyer Verlag.

Müller, Stefan. 2003. Mehrfache Vorfeldbesetzung. *Deutsche Sprache* 31(1). 29–62.

Müller, Stefan. 2004. Complex NPs, subjacency, and extraposition. *Snippets* 8. 10–11.

Müller, Stefan. 2007. Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus. In Gisela Zifonun & Werner Kallmeyer (eds.), *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*, 70–90. Berlin: Walter de Gruyter.

Müller, Stefan. 2013. Unifying everything: Some remarks on Simpler Syntax, Construction Grammar, Minimalism and HPSG. *Language* 89(4). 920–950.

Müller, Stefan. 2014. Kernigkeit: Anmerkungen zur Kern-Peripherie-Unterscheidung. In Antonio Machicao y Priemer, Andreas Nolda & Athina Sioupi (eds.), *Zwischen Kern und Peripherie*, 25–39. Berlin: de Gruyter.

Müller, Stefan. 2015. The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling* 3(1). 21–86.

Müller, Stefan. 2023a. *German clause structure: An analysis with special consideration of so-called multiple fronting*. Berlin: Revise and resubmit Language Science Press.

Müller, Stefan. 2023b. *Germanic syntax: A constraint-based view*. Berlin: Language Science Press.

Müller, Stefan. 2023c. *Grammatical theory: From Transformational Grammar to constraint-based approaches*. 5th edn. Berlin: Language Science Press.

Müller, Stefan. 2024. Anarchy, power, festschrifts, and universals. In Edward Gibson & Moshe Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*, ix–xxii. Berlin: Language Science Press.

Müller, Stefan, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.). 2024. *Head-Driven Phrase Structure Grammar: The handbook*. 2nd edn. Berlin: Language Science Press.

Newmeyer, Frederick J. 2005. *Possible and probable languages: A Generative perspective on linguistic typology*. Oxford: Oxford University Press.

Newport, Elissa L. 1990. Maturational constraints on language learning. *Cognitive Science* 14(1). 11–28.

Newport, Elissa L., Marc D. Hauser, Geertrui Spaepen & Richard N. Aslin. 2004. Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology* 49(2). 85–117.

Piantadosi, Steven T. 2024. Modern language models refute Chomsky's approach to language. In Edward Gibson & Moshe Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*, 353–414. Berlin: Language Science Press.

Pullum, Geoffrey K. 2024. Daniel Everett on Pirahã syntax. In Edward Gibson & Moshe Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*, 23–74. Berlin: Language Science Press.

Pullum, Geoffrey K. & Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1–2). 9–50.

Pulvermüller, Friedemann. 1999. Words in the brain's language. *Behavioral and Brain Sciences* 22(2). 253–279.

Pulvermüller, Friedemann, Bert Cappelle & Yury Shtyrov. 2013. Brain Basis of Meaning, Words, Constructions, and Grammar. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of Construction Grammar*, 397–416. Oxford: Oxford University Press.

Rice, Mabel. 1983. The role of television in language acquisition. *Developmental Review* 3(2). 211–224.

Richards, Marc. 2015. Minimalism. In Tibor Kiss & Artemis Alexiadou (eds.), *Syntax – theory and analysis: An international handbook*, vol. 2, 803–839. Berlin: Mouton de Gruyter.

Richter, Frank. 2021. Formal background. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*, 89–124. Berlin: Language Science Press.

Sag, Ivan A. 1997. English relative clause constructions. *Journal of Linguistics* 33(2). 431–483.

Scholz, Barbara C. & Geoffrey K. Pullum. 2002. Searching for arguments to support linguistic nativism. *The Linguistic Review* 19(1–2). 185–223.

Søgaard, Anders. 2023. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines* 33(1). 33–54.

Steels, Luc. 2003. Evolving grounded communication for robots. *Trends in Cognitive Science* 7(7). 308–312.

Strunk, Jan & Neal Snider. 2013. Subclausal locality constraints on relative clause extraposition. In Gert Webelhuth, Manfred Sailer & Heike Walker (eds.), *Rightward movement in a comparative perspective*, 99–143. Amsterdam: John Benjamins Publishing Co.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard & Julie C. Sedivy. 1996. Using eye movements to study spoken language comprehension: Evidence for visually mediated incremental interpretation. In Toshio Inui & James L. McClelland (eds.), *Information integration in perception and communication*, 457–478. Cambridge, MA: MIT Press.

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Paris: Librairie C. Klincksieck. Republished as *Elements of structural syntax*. Translated by Timothy Osborne and Sylvain Kahane. Amsterdam: John Benjamins Publishing Co., 2015.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, Michael. 2006. Construction Grammar for kids. *Constructions* Special Volume 1. 1–23.

Zhang, Kelly & Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In Tal Linzen, Grzegorz Chrupała & Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 359–361. Brussels, Belgium: Association for Computational Linguistics.