Ross D. Kristensen-McLachlan*, Pablo Contreras Kallens and Morten H. Christiansen

LLMs highlight the importance of interaction in human language learning

https://doi.org/10.1515/lingvan-2024-0254
Received December 23, 2024; accepted July 28, 2025; published online September 25, 2025

Abstract: Recent years have seen large language models (LLMs) achieve impressive performance with core linguistic abilities. This can be taken as a demonstration that, contrary to long-held assumptions about innate linguistic constraints, language can be learned through statistical learning from linguistic input alone. However, statistical language learning in these models differs from human learners in a crucial way: human language acquisition evolved not simply as passive absorption of linguistic input but is instead fundamentally interactive, guided by continuous feedback and social cues. Recent advances in LLM engineering have introduced an additional step in model training that utilizes more human-like feedback in what has come to be known as reinforcement learning from human feedback (RLHF). This procedure results in models which more closely mirror human linguistic behaviors – even reproducing characteristic human-like errors. We argue that the way RLHF changes the behavior of LLMs highlights how communicative interaction and socially informed feedback, in addition to input-driven statistical learning, can explain fundamental aspects of language learning. In particular, we take LLMs as models of "idealized statistical language learners" and RLHF as a form of "idealized language feedback", arguing that this perspective offers valuable insights into our understanding of human language development.

Keywords: large language models; interaction; language learning; feedback; learning models and linguistic theory

1 Introduction

Large language models (LLMs), while not full-fledged models of human language acquisition and use, provide clear evidence that a wide range of linguistic phenomena can be acquired through statistical learning from linguistic experience (Christiansen and Contreras Kallens 2022). LLMs consistently produce human-like grammatical novel outputs without the kind of built-in language-specific biases long assumed to be necessary for language learning (e.g., Chomsky 1965; Pinker 1994). Describing, testing, and understanding the limits and possibilities of their linguistic abilities is therefore an important goal for the cognitive science of language (Contreras Kallens et al. 2023). They provide one of the first working computational models with a more-or-less complete suite of linguistic abilities, something long thought to be out of reach for any neural network model (see, e.g., Pinker and Prince 1988).

Much contemporary discussion of LLMs within the language sciences has tended to adopt the (often implicit) view that LLMs are akin to passive "statistical sponges" (e.g., Chomsky et al. 2023). In many ways, this is broadly in agreement with the (again, often implicit) perspective of the dominant Chomskyan approach to human language acquisition (e.g., Chomsky 2005). Differences in learning mechanism notwithstanding, in both cases the central argument takes human experience of language to consist of passive exposure. However, there is a largely

***Corresponding author: Ross D. Kristensen-McLachlan**, Center for Humanities Computing, Aarhus University, Aarhus, Denmark; Center for Contemporary Cultures of Text, Aarhus University, Aarhus, Denmark; and Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Jens Chr. Skous Vej 4, 1483-522, Aarhus, Denmark, E-mail: rdkm@cc.au.dk. https://orcid.org/0000-0001-8714-1911
Pablo Contreras Kallens, Department of Language Science and Technology, Saarland University, Saarbrücken, Germany. https://orcid.org/0000-0002-3805-3488
Morten H. Christiansen, Center for Contemporary Cultures of Text, Aarhus University, Aarhus, Denmark; Interacting Minds Centre, Aarhus University, Aarhus, Denmark; and Department of Psychology, Cornell University, Ithaca, NY, USA. https://orcid.org/0000-0002-3850-0655

unnoticed difference between this passive image, prima facie more adequate for LLM training, and the type of experience with language that humans have during learning, which is fundamentally *interactive* (e.g., Christiansen and Chater 2022; Clark 1996; Goldstein and Schwade 2008; Pickering and Garrod 2004).

A recent crucial advance in LLM development has closed the gap on this qualitative difference. Previous generations of LLMs such as GPT-2 and the earliest versions of GPT-3 were trained in such a passive manner but most contemporary and state-of-the-art models, particularly consumer-facing ones, now undergo an additional training step known as reinforcement learning from human feedback (RLHF; Christiano et al. 2017; Ziegler et al. 2019). The goal of RLHF is to modify the behavior of the model to ensure that it performs specific tasks in ways which are considered more desirable by human users of the technology. This process of alignment is achieved primarily by collecting ranked preferences from human annotators and using them to reward the model for producing more preferable responses to user prompts and hence improved task performance specifically for potential human users.

In this position paper, we highlight the difference between passive and active learning in LLMs and argue that, beyond simply improved task performance, RLHF also has downstream effects on the linguistic abilities of LLMs in important ways. We draw on preliminary experimental results and argue that they not only reveal the limitations of the role of passive statistical learning in language acquisition; they also underscore the importance of interaction in the development of human language abilities. The incorporation of RLHF into LLM training thus opens new avenues for future research into explaining why human languages are the way they are.

2 LLMs as idealized language learners

It is clear that LLMs do not learn language in a way which maps directly onto human language learning. Contemporary models are exposed to volumes of input data far beyond what an individual human learner is exposed to, both in terms of volume and the range of different domains and contexts. For example, Meta's most recent flagship model Llama 3 is trained on 15.6 *trillion* tokens, over half of which appears to be natural language data (Dubey et al. 2024). Moreover, LLMs do not suffer from the same kinds of cognitive limitations experienced by human language learners, such as issues related to memory or retention (e.g., Christiansen and Chater 2016), with state-of-the-art LLMs retaining vast amounts of linguistic and nonlinguistic information present in their training data (Carlini et al. 2021; Chang et al. 2023; Chen et al. 2024; Tirumala et al. 2022). In fact, one of the key innovations of the transformer architecture is that the input is presented to the model in parallel and not sequentially (Vaswani et al. 2017), a complete departure from the inherently temporal nature of human linguistic experience (Elman 1990). All of this means that the linguistic model of an LLM implicit in its connection weights is not the model of any single language user as such, but instead represents a complex amalgam of different linguistic communities and cultures (Contreras Kallens and Christiansen 2025).

Nevertheless, not all possible models of the shared linguistic environment of a linguistic community are created equal. Indeed, apart from their impressive linguistic abilities, LLMs have been shown to be able to capture specifically human linguistic behavior better than other alternatives (e.g., Goldstein et al. 2022). We contend that this is because they acquire their linguistic ability specifically through STATISTICAL LEARNING. That is to say that these models begin with no predefined or a priori knowledge of language ahead of time but that this emerges over time through exposure to statistical regularities in natural language data.[1] Through repeated exposure to these regularities at scale, LLMs are ultimately able to learn how to produce language which, with few exceptions, is consistently grammatical or has error rates comparable to human language users (e.g., Contreras Kallens et al. 2023; Dou et al. 2022). If, as the last decades of research into human language acquisition suggests, humans also

1 Note, however, this does not mean that these models are blank slates but, rather, contain certain computational biases (Linzen and Baroni 2021). For example, when analyzing the internal representations of old-style connectionist networks, Christiansen and Chater (1999) found that the dynamics of *untrained* networks favored right-branching recursive structures over center-embedded ones. Crucially, though, these biases are domain-general in nature and thus unlike the kind of built-in language-specific properties previously assumed to be required for language acquisition (Chomsky 1965; Pinker 1994).

track these statistical regularities to learn language (e.g., Bogaerts et al. 2020; Isbilen and Christiansen 2022; Ruba et al. 2022; Saffran and Kirkham 2018), this is a key point of similarity between humans and LLMs (but for complementary arguments that statistical learning is not enough, see Frost et al. 2025). This relative human-likeness is further evidenced by recent work showing that LLMs are not as easily able to learn impossible languages compared to English as a control language (Kallini et al. 2024) and that LLMs can be used to perform metalinguistic reflection, such as eliciting grammaticality judgments (Begus et al. 2025; Hu et al. 2024; Hu and Levy 2023).

The core argument here is emphatically *not* that LLMs are a working model of exactly how language learning takes place in humans – far from it (see Birhane and McGann 2024). But the efficacy of models lies exactly in this abstraction from details: in their ability to provide insights on certain phenomena insofar as they resemble each other (Frigg and Hartmann 2025; Godfrey-Smith 2006; Morgan and Morrison 1999). In other words, in the context of language learning, we consider LLMs to be PROOFS-OF-CONCEPT (Portelance and Jasbi 2024). Given the way in which LLMs acquire core linguistic aptitude, they might be thought of as a kind of idealized language learner; and, again, insofar as human language learning can be said to resemble these mechanisms, studying LLMs can offer valuable insights for linguistic theory. In particular, as idealized language learners, unbounded by some, if not most, of the limitations that shape language processing in humans, LLMs can be used for testing and exploring the limits of statistical learning in language. And, as has been argued elsewhere by us and others, even when the link between language in humans and LLMs is kept in this "weak" form, the proficiency achieved by the latter upends key assumptions about the former (Contreras Kallens et al. 2023; Piantadosi 2024).

3 The role of feedback in language development

Notwithstanding the possible similarities between human learning and LLM training outlined above, there remains a fundamental difference. Whereas LLMs are passively exposed to cleaned and preprocessed text data to which they apply an algorithm optimized for modeling statistical regularities, humans learn language in socially situated, discursive, and interactive contexts with exposure to rich forms of FEEDBACK.

The role of feedback in language learning has been hotly debated in the language sciences for more than half a century. In the behaviorist tradition of the first half of the twentieth century, feedback was deemed important for language learning through interactions between stimuli and the responses they elicited (e.g., Skinner 1957). With the cognitive revolution following Chomsky's (1959) criticism of the behaviorist approach, the role of feedback in language learning was deemed minimal, if not nonexistent (e.g., Baker 1979; Marcus 1993). Instead, language learning was thought to be governed by a hypothesized innate capacity for language – often called a "universal grammar" (Chomsky 1965) – without which language learning would be impossible (Chomsky 2017; Jackendoff and Audring 2019; Yang et al. 2017).

In contrast to this perspective, a growing body of work under the umbrella of usage-based approaches eschewed the notion of an innate, specifically linguistic principles and/or computations in favor of more general cognitive mechanisms for statistical learning, abstraction, and generalization (e.g., Christiansen and Chater 2016; Goldberg 2019; Lieven 2014; Tomasello 2003). While much of this work relies entirely on observational data, particularly Bayesian approaches (see, e.g., Griffiths et al. 2024; Pearl and Goldwater 2016), usage-based research broadly construed has found evidence of the efficacy of feedback across different levels of linguistic structure, from pragmatic-discursive phenomena (Morgenstern et al. 2013) down to the level of phonetics and phonology (Goldstein and Schwade 2008; Kuhl 2007). The usefulness of feedback has also long been acknowledged in the second language learning literature (e.g., Leeman 2003; Long et al. 1998; Lyster and Ranta 1997; Muranoi 2000; Nassaji 2020).

Recently, feedback has also been employed in artificial language learning studies, which aim to study first and second language learning under carefully controlled laboratory conditions (e.g., Dale and Christiansen 2004; Jeuniaux et al. 2009; Monaghan et al. 2021). For example, in three experiments, Frinsel et al. (2024) showed that complex natural language structures, such as dative alternation (e.g., *Mary gave the book to John* vs. *Mary gave John the book*), required feedback to be learned successfully. Such learning was found both for negative feedback (when participants were told of incorrect responses) and positive feedback (when participants were told of

correct responses), and the effect was robust to both varying the probability of receiving feedback and intermingling negative and positive feedback.

Most of the work on feedback has focused on the possible role of explicit corrective feedback in language learning (e.g., Bohannon and Stanowicz 1988; Chouinard and Clark 2003). Such feedback includes corrections of specific phonological or syntactic errors (negative feedback), reformulations of incorrect sentences (negative feedback), or repetitions of correct sentences (positive feedback). However, there is another form of feedback that has received considerably less attention but which may be more important: communicative feedback (Nikolaus and Fourtassi 2023). Because children learn language through interaction with others (e.g., Casillas et al. 2024; Christiansen and Chater 2022; Elmlinger et al. 2023), the conversational behavior of their interlocutors can be used as implicit feedback on what the children just said. For example, adults provide implicit positive evidence when they nod or smile in response to something the child said (Tolins et al. 2017). And when adults ask for clarification of a misunderstood child utterance, they implicitly provide negative evidence to the child that they said something that was not quite right (Grosse et al. 2010). This kind of feedback is ubiquitous in everyday language use (Dideriksen et al. 2023) and may thus provide a rich source of feedback to scaffold children's language learning (Nikolaus and Fourtassi 2023).

4 RLHF as idealized feedback

As the name suggests, RLHF is a form of reinforcement learning (RL), which has a long, rich history as an approach to machine learning (Sutton and Barto 2018). It is designed primarily to explore solutions to complex tasks with goals that are difficult to define or specify, making it particularly well suited for the kinds of tasks regularly encountered in computer vision and, more importantly for our purposes, natural language processing (Christiano et al. 2017; Ziegler et al. 2019). The mathematics underlying RL goes beyond the scope of this paper but the essentials can be grasped quite intuitively, since RL simply involves agents learning how to act appropriately in response to external stimuli. An RL agent utilizes a POLICY which specifies how to select actions given some STATE derived from these stimuli, with different states being weighted by a REWARD FUNCTION which prioritizes preferred states.

In other words, the goal of RL is to create agents which can learn more effectively to respond and act based on environmental feedback. Traditional approaches to this task generally had a predefined reward function, with the model learning the optimum policy given the data (for a full survey, see Kaufmann et al. 2024). However, this approach very quickly runs into an obvious problem, which is that explicitly defining a reward function can be extremely challenging for sufficiently complex tasks. For example: what is the goal of language learning? This is a question for linguistic theory and one which is likely to generate as many different responses as there are competing theories of language. Likewise, even the most well-defined reward functions may have unexpected outcomes as a result of blind spots and oversights. Given the seeming intractability of these questions, much recent research in RL has instead gravitated towards REWARD LEARNING, where a REWARD MODEL is learned in parallel to the policy.

RLHF is a particularly successful application of this kind of reward learning where both the reward model and policy are learned based on explicit feedback from humans. For most contemporary LLMs, RLHF is achieved through human annotators ranking multiple outputs from LLMs trained directly on natural language data (merely optimizing predictive success), with preferred outputs ranked more highly (Ouyang et al. 2022). This particular training regime has been so successful that it now forms the basis of nearly all of the most highly performant consumer-facing LLMs (Casper et al. 2023; Kaufmann et al. 2024). Given the kinds of goal-oriented motivation common to machine learning research, RLHF has been widely adopted in contemporary natural language processing primarily because systems using ranked human preferences tend to result in better performance on downstream tasks. In the context of interactive LLMs or "chatbots" such as ChatGPT, this feedback has the additional effect of creating models which generate language more similar to the kinds of outputs users want to see.

As such, the role of RLHF in the case of LLMs can be seen as analogous to interactive feedback in human language acquisition, ensuring that an individual's learned model of language more closely matches the expectations of the language community. Indeed, RL has been studied as a paradigm of interactive language learning, such as in vocal motor development (Warlaumont et al. 2013), dialogue modeling (Khalid et al. 2020), and

elements of Gricean and relevance-theoretic pragmatic phenomena (Sumers et al. 2024). RL has also been used to model key aspects of language evolution, such as the emergence of representational alignment (Kouwenhoven et al. 2024). Since LLMs passively exposed to natural language data can be understood as idealized language learners, we propose that RLHF can be understood analogously as a kind of IDEALIZED LANGUAGE FEEDBACK, promoting alignment with broader norms around language use that go beyond its sole statistical structure. But how far does this analogy go? Beyond simply improving performance on technical benchmarks, how much – if at all – does RLHF modify the actual linguistic output produced by LLMs?

We can begin to approach these questions experimentally by drawing on human behavioral data from previous research and comparing these human responses to LLMs trained both with and without RLHF. A simple, compelling example of this can be demonstrated by drawing on data taken from previous experiments comparing language models with human behavioral data. Linzen et al. (2016) tested comparatively simple recurrent neural networks and found them to significantly outperform humans when it comes to accurately learning long-distance dependencies in a subject–verb agreement task in English with either *is* or *are* as the target verb and different types of distractors in the sentence. Using their data, we designed and preregistered a series of simple experiments to determine the effect of RLHF by prompting both an LLM acting as a "passive sponge" (davinci, or "vanilla") and a model trained with RLHF (text-davinci-003, or "chat"). Both models are of OpenAI's GPT-3 family, with the main difference between them seemingly being the introduction of RLHF for text-davinci-003.

Our data was generated using the OpenAI Completions API for Python (https://platform.openai.com/docs/api-reference/completions). For each experimental item from Linzen et al. (2016), we zero-shot prompted each model to generate a completion for that sentence fragment, with the prompt given to the models mirroring the instructions shown to the human subjects as closely as possible, including that the correct answer is either *is* or *are*. We calculated the probabilities of the models continuing a given sequence with *is* or *are* immediately after the input sequence. The temperature was set to 0, which means that the model's predictions were deterministic, so each sentence was completed only once per condition. There are two manipulations for each sentence. First, the number of each of the nouns in subject and object positions is varied (S = singular; P = plural), as illustrated in examples (1)–(4):

(1) *The dog who ate the cake* ***is/are*** (S/S)

(2) *The dogs who ate the cake* ***is/are*** (P/S)

(3) *The dog who ate the cakes* ***is/are*** (S/P)

(4) *The dogs who ate the cakes* ***is/are*** (P/P)

And second, the type of clause is also varied (R = relative; P = prepositional), as seen in (5) and (6):

(5) *The dog who ate the cakes* ***is/are*** (S/P/R)

(6) *The dog with the cat* ***is/are*** (S/S/P)

The sensitivity of agreement processes in English to these kinds of attractor effects is well known and widely attested in the psycholinguistic literature, both in natural production of speech (Bock and Miller 1991) and in comprehension tasks (Pearlmutter et al. 1999). Linzen et al. (2016) found that humans were prone to making mistakes when there are mismatches of number in the nouns, particularly when the subject is singular and the intervening noun is plural, both in relative and prepositional clauses, with a stronger effect for the latter.

Figure 1 shows the results of this experiment. The different spokes on the radar charts represent the different types of sentences with which humans were tested, varying both number and type of clause. The overlaid polygons show in blue the performance of humans in Linzen et al. (2016) and in yellow and red the performance of the vanilla and chat models, respectively. Performance has been normalized to the 0–1 range, with 1 being the type of sentence with the higher performance. For humans, these are the filler sentences, with all but P/S/R (0.67), S/P/R (0.21), and S/P/P (0) around the 0.85 performance mark. For the vanilla davinci model on the left, the attractors cause no problem for the LLM, which can predict the correct verb in this context in various conditions: fillers are still the highest, with S/P/P (0.95), S/S/P (0.95), and P/P/P (0.85) close behind it. In contrast to humans, however, the vanilla model struggles with relative clauses even with no attractor, with low proportional accuracy

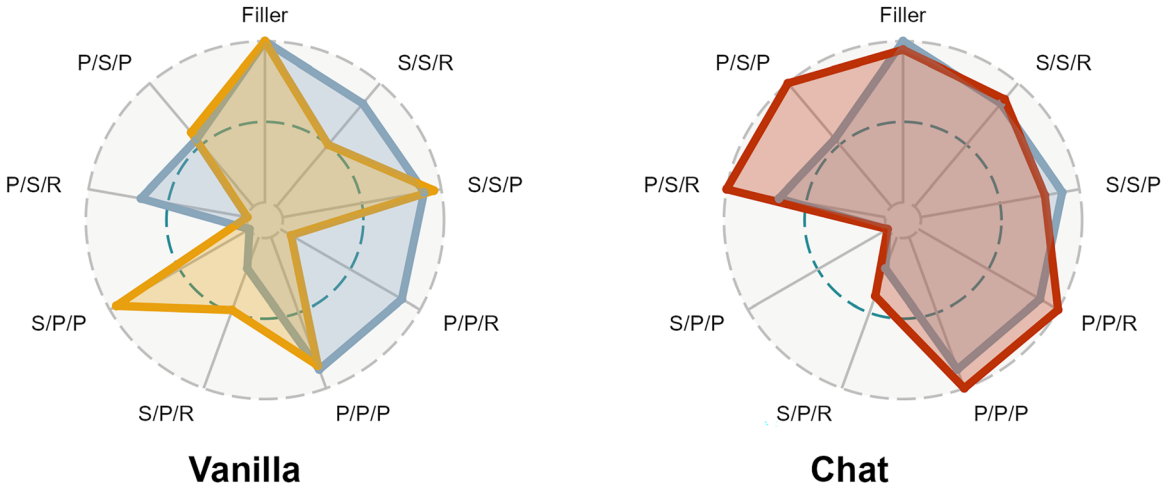

Figure 1: Relative performance for the base davinci model (or "vanilla", in yellow) compared to the RLHF trained text-davinci-003 (or "chat", in red). Human performance is shown in blue. Each sentence was completed once with a model temperature of 0. The overall performance on each sentence type was min-max normalized with the best performance coded a 1 and the worst performance as 0. The dashed line marks 0.5. Each spoke represents a different condition. The first letter is the number of the subject (S for singular, P for plural), the second letter is the number of the attractor noun (S for singular, P for plural), and the third letter represents the type of clause (P for prepositional, R for relative) (adapted from Contreras Kallens et al. 2025).

in S/S/R (0.5), P/P/R (0.08), S/P/R (0.48), and P/S/R (0). The Pearson correlation between the normalized scores between humans and the vanilla model is nonexistent, at −0.04 (t = −0.11, df = 7, p = 0.92).

By contrast, the model trained with the additional RLHF step displays a more human-like performance on most conditions. The maximum scores are in the filler, P/P/R, P/P/P, and in the clauses with singular attractors, P/S/R and P/S/P. What is more intriguing, however, is that the worst performance is on sentences with plural attractors, S/P/R (0.39) and S/P/P (0), as illustrated in examples (7) and (8):

(7) *The dog who ate the cakes* ***is/are*** (S/P/R)

(8) *The dog with the cats* ***is/are*** (S/P/P)

This similitude is in marked contrast to the vanilla model, with a Pearson correlation with the human scores of 0.86 (t = 4.54, df = 7, p = 0.003). It is worth remarking that, in this case, the RLHF training was detrimental to absolute performance in the task, as the accuracy in S/P/P sentences was 0.97 for the vanilla model compared to 0.28 for the chat model.

The crucial implication to be drawn from these results, although preliminary, is that the LLM supplemented with human feedback produces more human-like outputs *including reproducing the kinds of errors commonly made by humans*. This seems counterintuitive, given that the additional feedback results in models which perform worse from a purely objective grammatical perspective and instead appear to mimic human patterns of incorrect performance. It is also surprising, as imitating human grammatical errors is not part of the explicit goals pursued by the RLHF procedure. Moreover, it is unlikely that feedback is the reason for similar mistakes in humans, for whom agreement processing is a complex psycholinguistic process (Kandel and Colin 2022; Kandel et al. 2022). One possible explanation for what is happening is that agreement errors are already present in the training data to which the vanilla model has been exposed, but that they are dispreferred and hence not regularly generated by the vanilla model. Depending on the annotation guidelines and individual preferences, human annotators may unconsciously rank more highly responses with agreement mismatches. The point here then is that RLHF does not create novel behavior but instead amplifies latent potential, steering the LLM towards the preference of a given community-level linguistic environment (Contreras Kallens and Christiansen 2025).[2]

2 We wish to thank an anonymous reviewer for drawing our attention to this suggestion.

Nevertheless, our initial experiments demonstrate that the introduction of ranked human preferences results in models that behave linguistically in different ways from those simply exposed to large volumes of natural language data. Perhaps even more striking is that, since RLHF alone is responsible for weight updates in the chat-based model, the statistical regularities learned from exposure to trillions of tokens of natural language appear to have been partially overridden by exposure to comparatively little in the way of feedback. Far from being passive statistical sponges that recapitulate the statistics of their massive input, then, contemporary LLMs actively learn to optimize their language production based on the feedback they receive from their environment. That this happens at all is a nontrivial result with potentially far-reaching implications for theories of language acquisition and use as well as offering insight into more applied efforts to control and improve model training. It therefore seems that careful experimentation designed to probe the limits and possibilities inherent to LLMs as models of human language is a vital testbed for linguistic theory and the wider language sciences.

5 Conclusions

In this position paper, as elsewhere, we have argued that LLMs provide working proof that core linguistic competencies can be acquired through statistical learning from distributional properties of natural language. That this is possible at all suggests that humans can also learn to process a grammar purely through statistical learning, without the need for a particular faculty dedicated to language acquisition. However, empirical evidence suggests that statistical learning is not the be-all and end-all when it comes to human language acquisition, with social, communicative, and interactive feedback playing a foundational role in the development of human language in addition to the ability to purely track statistical patterns. Our experiment suggests that feedback plays a similarly crucial role in the way contemporary LLMs learn to process and generate natural language, even causing models to produce more errors in order to be more human-like.

What does this mean for linguistics and the cognitive science of language? We do not wish to argue that human language learning or use occurs in exactly the same ways that LLMs learn and use language. Nor do we wish to suggest that RLHF is fully comparable with the rich cultural, social, and interactive linguistic feedback to which humans are exposed. Instead, we contend that, just as LLMs can be understood as idealized language learners, RLHF can be understood as a model of some aspects of language feedback. Although "implemented" in widely different ways, both humans and LLMs learn the statistical patterns from exposure to language data, and both seem to require additional alignment, humans to their linguistic communities and LLMs to the relevant community in which they are to be used (typically contemporary American English). In this way, LLMs seem to capture meaningful dimensions of a complex and adaptive cultural system which both guides and is shaped by learning and communication through processes that resemble some relevant aspects of human language learning (Contreras Kallens and Christiansen 2025). This system, human language, has evolved for the purpose of interactive use in conversation (Christiansen and Chater 2022; Roberts and Levinson 2017; Tomasello 1999). Insofar as contemporary LLMs highlight this fact, they continue to provide compelling computational tools to understand why languages are the way they are.

References

Baker, Carl Leroy. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10. 533–581.

Begus, Gasper, Maksymilian Dabkowski & Ryan Rhodes. 2025. Large linguistic models: Investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence* 1–15.

Birhane, Abeba & Merek McGann. 2024. Large models of what? Mistaking engineering achievements for human linguistic agency. *Language Sciences* 106. https://doi.org/10.1016/j.langsci.2024.101672.

Bock, Kathryn & Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology* 23(1). 45–93.

Bogaerts, Louisa, Ram Frost & Morten H. Christiansen. 2020. Integrating statistical learning into cognitive science. *Journal of Memory and Language* 115. https://doi.org/10.1016/j.jml.2020.104167.

Bohannon, John N. & Laura B. Stanowicz. 1988. The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology* 24(5). 684–689.

Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song Úlfar Erlingsson, Alina Oprea & Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX security 21)*, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting (accessed 24 August 2025).

Casillas, Marisa, Naja Ferjan Ramírez, Victoria Leong & Romeo Rachel. 2024. Becoming a conversationalist: Questions, challenges, and new directions in the study of child interactional development. *Infant Behavior and Development* 76. https://doi.org/10.1016/j.infbeh.2024.101956.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, J'er'emy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro J Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco di Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh & Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint. https://doi.org/10.48550/arXiv.2307.15217.

Chang, Kent, Mackenzie Cramer, Sandeep Soni & David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 conference on empirical methods in natural language processing (EMNLP)*, 7312–7327. Singapore: Association for Computational Linguistics.

Chen, Bowen, Namgi Han & Yusuke Miyao. 2024. A multi-perspective analysis of memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 11190–11209. Miami, FL: Association for Computational Linguistics.

Chomsky, Noam. 1959. A review of B. F. Skinner's verbal behavior. *Language* 35. 26–57.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36(1). 1–22.

Chomsky, Noam. 2017. The language capacity: Architecture and evolution. *Psychonomic Bulletin & Review* 24. 200–203.

Chomsky, Noam, Ian Roberts & Watumull Jeffrey. 2023. The false promise of Chat-GPT. *New York Times* 8 March. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Chouinard, Michelle M. & Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language* 30(3). 637–669.

Christiano, Paul F., Leike Jan, Tom Brown, Miljan Martic, Shane Legg & Dario Amodei. 2017. Deep reinforcement learning from human preferences. arXiv preprint https://doi.org/10.48550/arXiv.1706.03741.

Christiansen, Morten H. & Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23. 157–205.

Christiansen, Morten H. & Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* 39. e62.

Christiansen, Morten H. & Nick Chater. 2022. *The language game: How improvisation created language and changed the world*. New York, NY: Basic Books.

Christiansen, Morten H. & Pablo Contreras Kallens. 2022. AI is changing scientists' understanding of language learning – And raising questions about an innate grammar. *The Conversation* 19 October. https://theconversation.com/ai-is-changing-scientists-understanding-of-language-learning-and-raising-questions-about-an-innate-grammar-190594.

Clark, Herbert H. 1996. *Using language*. Cambridge: Cambridge University Press.

Contreras Kallens, Pablo & Morten H. Christiansen. 2025. Distributional semantics: Meaning through culture and interaction. *Topics in Cognitive Science* 17(3). 739–769.

Contreras Kallens, Pablo, Ross D. Kristensen-McLachla & Morten H. Christiansen. 2025. Human feedback makes large language models more human-like. Unpublished manuscript.

Contreras Kallens, Pablo, Ross D. Kristensen-McLachlan & Morten H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science* 47(3). 1–6.

Dale, Rick & Morten H. Christiansen. 2004. Active and passive statistical learning: Exploring the role of feedback in artificial grammar learning and language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26. https://escholarship.org/uc/item/4k77621p.

Dideriksen, Christina, Morten H. Christiansen, Kristian Tylén, Mark Dingemanse & Riccardo Fusaroli. 2023. Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General* 152(3). 864–889.

Dou, Yao, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith & Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers)*, 7250–7274. Dublin: Association for Computational Linguistics.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur'elien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cris-tian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song,

Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin R. Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen-ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Niko-lay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron-nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa-hana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Vir-ginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit-ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto de Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthias Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Mun-ish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navy-ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pe-dro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Kumar Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang & Zhiwei Zhao. 2024. The llama 3 herd of models. arXiv preprint. https://doi.org/10.48550/arXiv.2407.21783.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14(2). 179–211.
Elmlinger, Steven L., Jennifer A. Schwade, Laura Vollmer & Michael H. Goldstein. 2023. Learning how to learn from social feedback: The origins of early vocal development. *Developmental Science* 36. https://doi.org/10.1111/desc.13296.
Frigg, Roman & Stephan Hartmann. 2025. Models in science. In E. N. Zalta (ed.), *The stanford encyclopedia of philosophy (summer 2025)*. https://plato.stanford.edu/archives/sum2025/entries/models-science.
Frinsel, Felicity F., Trecca Fabio & Morten H. Christiansen. 2024. The role of feedback in the statistical learning of language-like regularities. *Cognitive Science* 48(3). 1–30.
Frost, Ram, Bogaerts Louisa, G. Samuel. Arthur, James S. Magnuson, Lori L. Holt & Morten H. Christiansen. 2025. Statistical learning subserves a higher purpose: Novelty detection in an information foraging system. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000547.
Godfrey-Smith, Peter. 2006. The strategy of model-based science. *Biology and Philosophy* 21(5). 725–740.
Goldberg, Adele. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton, NJ: Princeton University Press.
Goldstein, Michael H. & Jennifer A. Schwade. 2008. Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science* 19(5). 515–523.
Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky & Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* 25(3). 369–380.
Griffiths, Thomas L., Nick Chater & Joshua Tenenbaum (eds.). 2024. *Bayesian models of cognition: Reverse engineering the mind*. Cambridge, MA: MIT Press.
Grosse, Gerlind, Tanya Behne, Malinda Carpenter & Michael Tomasello. 2010. Infants communicate in order to be understood. *Developmental Psychology* 46. 1710–1722.
Hu, Jennifer & Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 5040–5060. Singapore: Association for Computational Linguistics.
Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova & Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences USA* 121(36). https://doi.org/10.1073/pnas.2400917121.
Isbilen, Erin S. & Morten H. Christiansen. 2022. Statistical learning of language: A meta-analysis into 25 years of research. *Cognitive Science* 46(9). https://doi.org/10.1111/cogs.13198.
Jackendoff, Ray & Jenny Audring. 2019. The parallel architecture. In András Kertész, Edith Moravcsik & Csilla Rákosi (eds.), *Current approaches to syntax: A comparative handbook*, 215–240. Berlin: De Gruyter Mouton.
Jeuniaux, Patrick, Rick Dale & Max M. Louwerse. 2009. The role of feedback in learning form meaning mappings. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 31. https://escholarship.org/uc/item/53k5m0j2.
Kallini, Julie, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald & Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers)*, 14691–14714. Bangkok: Association for Computational Linguistics.
Kandel, Margaret & Phillips Colin. 2022. Number attraction in verb and anaphor production. *Journal of Memory and Language* 127. https://doi.org/10.1016/j.jml.2022.104370.
Kandel, Margaret, Cassidy Rae Wyatt & Colin Phillips. 2022. Agreement attraction error and timing profiles in continuous speech. *Glossa Psycholinguistics* 1(1). 1–46.
Kaufmann, Timo, Weng Paul, Viktor Bengs & Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. arXiv preprint. https://doi.org/10.48550/arXiv.2312.14925.
Khalid, Baber, Malihe Alikhani & Matthew Stone. 2020. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4417–4428. Barcelona: International Committee on Computational Linguistics.
Kouwenhoven, Tom, Max Peeperkorn, Bram Van Dijk & Tessa Verhoef. 2024. The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 57–71. Bangkok: Association for Computational Linguistics.
Kuhl, Patricia K. 2007. Is speech learning "gated" by the social brain? *Developmental Science* 10(1). 110–120.
Leeman, Jennifer. 2003. Recasts and second language development: Beyond negative evidence. *Studies in Second Language Acquisition* 25(1). 37–63.
Lieven, Elena. 2014. First language development: A usage-based perspective on past and current research. *Journal of Child Language* 40(Suppl. 1). 48–63.
Linzen, Tal, Emmanuel Dupoux & Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4. 521–535.
Linzen, Tal & Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics* 7(1). 195–212.

Long, Michael, Shunji Inagaki & Lourdes Ortega. 1998. The role of implicit negative feed-back in SLA: Models and recasts in Japanese and Spanish. *The Modern Language Journal* 82. 357–370.
Lyster, Ryan & Leila Ranta. 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition* 20. 37–66.
Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition* 46(1). 53–85.
Monaghan, Padraic, Simón Ruiz & Patrick Rebuschat. 2021. The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research* 37. 261–289.
Morgan, Mary S. & Margaret Morrison (eds.). 1999. *Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.
Morgenstern, Aliyah, Marie Leroy-Collombel & Stéphanie Caët. 2013. Self- and other-repairs in child–adult interaction at the intersection of pragmatic abilities and language acquisition. *Journal of Pragmatics* 56. 151–167.
Muranoi, Hitoshi. 2000. Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning* 50. 617–673.
Nassaji, Hossein. 2020. Assessing the effectiveness of interactional feedback for L2 acquisition: Issues and challenges. *Language Teaching* 53. 3–28.
Nikolaus, Mitja & Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology* 68. 1–11.
Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Hilton Jacob, Kelton Fraser, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Christiano Paul, Leike Jan & Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, article 2011, 27730–27744. Red Hook, NY: Curran Associates .
Pearl, Lisa & Sharon Goldwater. 2016. Statistical learning, inductive bias, and Bayesian inference in language acquisition. In Jeffrey Lidz, William Snyder & Joe Pater (eds.), *The Oxford handbook of developmental linguistics*, 664–695. Oxford: Oxford University Press.
Pearlmutter, Neal J., Susan M. Garnsey & Kathryn Bock. 1999. Agreement processes in sentence comprehension. *Journal of Memory and Language* 41(3). 427–456.
Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1–2). 73–193.
Piantadosi, Steven T. 2024. Modern language models refute Chomsky's approach to language. In Edward Gibson & Moshe Polanyi (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*, 353–414. Berlin: Language Science Press.
Pickering, Martin J. & Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2). 169–226.
Pinker, Steven. 1994. *The language instinct: How the mind creates language*. New York, NY: William Morrow.
Portelance, Eva & Masoud Jasbi. 2024. The roles of neural networks in language acquisition. *Language and Linguistics Compass* 18(6). https://doi.org/10.1111/lnc3.70001.
Roberts, Seán G. & Stephen C. Levinson. 2017. Conversation, cognition and cultural evolution: A model of the cultural evolution of word order through pressures imposed from turn taking in conversation. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems* 18(3). 402–442.
Ruba, Ashley L., Seth D. Pollak & Jenny R. Saffran. 2022. Acquiring complex communicative systems: Statistical learning of language and emotion. *Topics in Cognitive Science* 14(3). 432–450.
Saffran, Jenny R. & Natasha Z. Kirkham. 2018. Infant statistical learning. *Annual Review of Psychology* 69. 181–203.
Skinner, Burrhus F. 1957. *Verbal behavior*. New York, NY: Appleton-Century-Crofts.
Sumers, Theodore R., Mark K. Ho, Thomas L. Griffiths & Robert D. Hawkins. 2024. Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review* 131(1). 194–230.
Sutton, Richard S. & Andrew G. Barto. 2018. *Reinforcement learning: An introduction*, 2nd edn. Cambridge, MA: MIT Press.
Tirumala, Kushal, Aram H. Markosyan, Luke Zettlemoyer & Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, 38274–38290. Red Hook, NY: Curran Associates.
Tolins, Jackson, Neda Namiranian, Nameera Akhtar & Jean E. Fox Tree. 2017. The role of addressee backchannels and conversational grounding in vicarious word learning in four-year-olds. *First Language* 37(6). 648–671.
Tomasello, Michael. 1999. *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. arXiv preprint. https://doi.org/10.48550/arXiv.1706.03762.
Warlaumont, Anne S., Gert Westermann, Eugene H. Buder & D. Kimbrough Oller. 2013. Prespeech motor learning in a neural network using reinforcement. *Neural Networks* 38. 64–75.
Yang, Charles, Stephen Crain, Robert C. Berwick, Noam Chomsky & Johan J. Bolhuis. 2017. The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews* 81(Pt B). 103–119.
Ziegler, Daniel M., Stiennon Nisan, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Christiano Paul & Geoffrey Irving. 2019. Fine-tuning language models from human preferences. arXiv preprint. https://doi.org/10.48550/arXiv.1909.08593.