frontiers in Psychology

MINI REVIEW
published: 23 March 2022
doi: 10.3389/fpsyg.2022.741321

Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition?

Ludovica Pannitto[1] *and Aurelie Herbelot*[1,2]*

[1] CIMeC - Centre for Mind and Brain Sciences, University of Trento, Trento, Italy, [2] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

It has been shown that Recurrent Artificial Neural Networks automatically acquire some grammatical knowledge in the course of performing linguistic prediction tasks. The extent to which such networks can actually learn grammar is still an object of investigation. However, being mostly data-driven, they provide a natural testbed for usage-based theories of language acquisition. This mini-review gives an overview of the state of the field, focusing on the influence of the theoretical framework in the interpretation of results.

Keywords: recurrent neural networks, grammar, usage-based linguistics, language acquisition, construction grammar

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Alex Warstadt,
New York University, United States
Alessandra Falzone,
University of Messina, Italy

***Correspondence:**
Aurelie Herbelot
aurelie.herbelot@unitn.it

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *14 July 2021*
Accepted: *25 February 2022*
Published: *23 March 2022*

Citation:
Pannitto L and Herbelot A (2022) Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition? Front. Psychol. 13:741321. doi: 10.3389/fpsyg.2022.741321

1. INTRODUCTION

Artificial Neural Networks (ANNs), and in particular recurrent architectures such as Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997), have consistently demonstrated great capabilities in the area of language modeling, generating sentences with credible surface patterns and showing promising performance when tested on very specific grammatical abilities (Gulordava et al., 2018; Linzen and Baroni, 2021), without requiring any prior bias towards the syntactic structure of natural languages. From a theoretical point of view, however, published results sometimes appear inconsistent, and overall inconclusive. The present survey suggests however that results should be interpreted in the light of various theoretical frameworks if they are to be fully understood. To illustrate this, it approaches the literature from the point of view of usage-based theories of acquisition, which are naturally suited to the behaviorist setting implemented by language modeling techniques.

2. USAGE-BASED THEORIES OF GRAMMAR ACQUISITION

Taking a coarse-grained perspective on usage-based theories of language acquisition, we can pinpoint three main standpoints that are relevant to language modeling with ANNs.

First and foremost, behaviorist theories argue for a systemic vision where general-purpose memory and cognitive mechanisms account for the emergence of linguistic abilities (Tomasello, 2003; Goldberg, 2006; Christiansen and Chater, 2016; Cornish et al., 2017). That is, they stand against the idea that explicit, *innate* biases should be required in the acquisition device.

Secondly, usage-based theories argue for a tight relation between *input* and learned representations in the course of acquisition (Jackendoff, 2002; Boyd and Goldberg, 2009). This is based on results that indicate that infants understand and manipulate input signals in sophisticated ways: their ability to analyze stream-like signals like language is well explored in the statistical learning literature (Gómez and Gerken, 2000; Romberg and Saffran, 2010; Christiansen, 2019), and the shape of the input itself has been explained by its relation to basic cognitive processes (Christiansen and Chater, 2015; Cornish et al., 2017). Word segmentation for instance is accomplished by 8-month old infants, relying purely on statistical relationships between neighboring speech sounds, and with very limited exposure (Saffran et al., 1996). Such limited input is also enough for one-year-olds to acquire specific grammatical information, thus discriminating new grammatical strings from those that show string-internal violations (Gomez and Gerken, 1999).

Thirdly, gradedness of grammatical notions is a central aspect in usage-based theories. Cognitive theories tend to blur hard boundaries, e.g. when it comes to the structure of categories (Barsalou, 1987), the content of semantic knowledge (Elman, 2009; McRae and Matsuki, 2009) or the distinction between lexically filled and pattern-like instances (Goldberg, 2006).

Artificial statistical models seem an ideal toolbox to test the above claims. They can be built without hard-coded linguistic biases and they can be fed different types of input to investigate their effect on the acquisition process. Moreover, both their behavior and internal state can be analyzed in various ways. Lakretz et al. (2019) take a physiological approach investigating how, with no explicit bias, specific neurons specialize in detecting and memorizing syntactic structures. Giulianelli et al. (2018) propose instead a diagnostic downstream classifier to evaluate representations of number agreement.

The rest of this survey approaches the literature in the light of the three aspects of usage-based frameworks mentioned above, discussing to what extent the theory fits both implementation and results.

3. NEURAL LANGUAGE MODELS AND LANGUAGE DEVELOPMENT

The comparison between artificial language models and human language development starts at a fundamental mechanism: prediction. Predictive functions are considered highly relevant to language processing (Pickering and Garrod, 2013; Ramscar et al., 2013) and have received particular attention from theories that posit a direct relation between the shape of the received input and the organization of grammar (Ramscar et al., 2013; Fazekas et al., 2020). Consequently, (artificial) predictive models should be ideally suited to test related hypotheses.

While prediction is a shared mechanisms among neural architectures, different models have been specialized for different tasks, leveraging prediction in various ways. The task most relevant to this survey is known as Language Modeling (LM): networks are trained to *predict* the next word (or character) given the previous sequence. Language Modeling encodes language competence only partially, leaving aside aspects such as interaction, grounding or event knowledge, which are crucial to human linguistic abilities. Nevertheless, it lets us test to what extent grammar can be learned from a pure and linear linguistic signal.

Recurrent Neural Networks (RNNs), and more specifically the "Long Short-Term Memory network" or **LSTM** (see **Figure 1** for a brief description), are among the most common architectures and the ones with the longest history in Language Modeling. In LSTMs, contextual information is maintained from one prediction step to the next. The output of the network at time t thus depends on a subset of the inputs fed to the network across a time window. The LSTM learns to regulate its attention over this time window, deciding what to remember and what to forget in the input.

LSTMs are a useful framework to compare learning in a purely predictive setting and an innately biased model. Expectedly, LSTMs that carry explicit syntactic bias [e.g. Recurrent Neural Network Grammars, Dyer et al. (2016); Kuncoro et al. (2017)] and specifically highlight the benefits of top-down parsing as an anticipatory model (Kuncoro et al., 2018) tend to perform better in experiments. But the question asked by usage-based theories is to what extent such hard-coded biases could be learned from language exposure only. A prime example of the pure prediction approach can be found in Gulordava et al. (2018): a vanilla LSTM is trained on a Language Modeling task, under the argument that the predictive mechanism is sufficient for the network to predict long-distance number agreement. The authors conclude that "LM-trained RNNs can construct abstract grammatical representations." In a more ambivalent study, Arehalli and Linzen (2020) consider how real-time human comprehension and production do not always follow the general grammatical constraint of subject-verb agreement, due to a variety of possible syntactic or semantic factors. They replicate six experiments from the agreement attraction literature using LSTMs as subjects, and find that the model, despite its relatively simple structure, captures human behavior in at least three of them. The authors argue that those phenomena can be regarded as emerging from domain-general processing mechanisms, while also conceding that additional mechanisms might be required to model others.

Notably, LSTMs also process the linguistic signal incrementally, and can be trained on relatively small amounts of data, comparable to the quantities that children are exposed to during the acquisition years (Hart et al., 1997). While this does not make LSTMs plausible models of human cognition, it makes them good benchmarks for building and verifying a range of psycholinguistic hypotheses around incremental processing and the poverty of the stimulus. This feature is especially important to test usage-based ideas that the statistical distribution of child-directed language explains how children acquire constructions in spite of the limited input they receive (see Section 4).

More recently, a new class of models has emerged and shown excellent performance in generating natural language (i.e.,

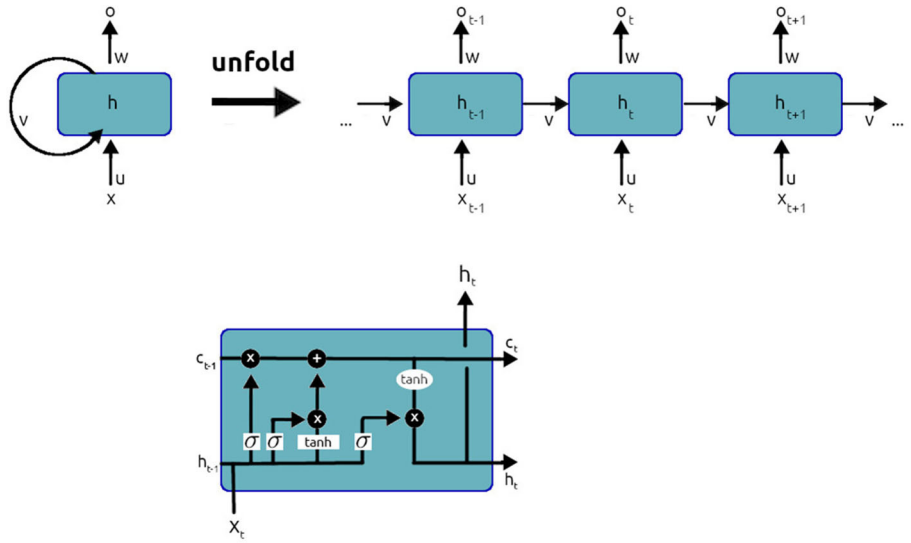

FIGURE 1 | LSTM networks are capable of keeping track of long-term dependencies. As recurrent neural networks (upper layer of the figure), they present a chain-like structure: at each time step t, the network's output is computed based on both the input of time $t(x_t)$ and the network's state at time $t-1(h_{t-1})$. As opposed to a simple recurrent cell, an LSTM cell (lower layer of the figure) has the ability to regulate how the two kinds of information (input and previous state) are weighted towards the computation of the output. The first gate, the forget gate, evaluates C_{t-1} (a representation of the previous state different from h_{t-1}) against x_t and learns what information to keep from previous steps, including it in a vector f_t. Next, a candidate value for the current state $\hat{C}_t$ is computed along with the input gate vector i_t that weighs how much of the input will contribute to the current state. Finally, the state of the cell C_t is computed by weighting C_{t-1} with the forget gate vector f_t and the at $\hat{C}_t$ with the input vector i_t. h_t is then computed from C_t. A complete and easy to read guide to LSTMs can be found at https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Transformer models Vaswani et al., 2017, TLMs) and have in fact been shown to learn structural biases from raw input data (Warstadt and Bowman, 2020). Some psycholinguistic informed approaches have emerged around the architecture. Related the question of acquisition, Warstadt et al. (2020a) and Hu et al. (2020) have compared a range of models, including LSTMs and transformers, on different sizes of corpora. While the amount of training input clearly benefits system performance, Hu et al. (2020) also conclude that the specific hard-coded architecture of a model is more important than data size in yielding correct syntactic knowledge. Their training data is however not characteristic of child-directed input. In contrast, Huebner et al. (2021) focus on training a TLM on developmentally plausible input, matched in quantity and quality to what children are exposed to. The authors also introduce a novel test suite compatible with child-directed language requirements, such as a reduced vocabulary. Their results show that both features of the input and hyperparameters setting are highly relevant for the acquisition process.

While TLMs seem to be a promising new avenue for researchers, they require very large amounts of data to be trained and exhibit a real preference for linguistic generalization, as opposed to surface patterns (Warstadt et al., 2020b). It is also still unclear whether such networks truly generalize or simply memorize patterns they have encountered, leveraging their extremely large size (Kharitonov et al., 2021).

4. THE ROLE OF INPUT

While widely debated in linguistic research, the effect of *input* on learning has received less attention in computational studies, due to the lack of availability of diverse and realistic input data. This aspect is however a pillar of usage-based theories, and can help make sense of various studies that report seemingly inconsistent results across different input data.

Starting with the issue of input size, experiments such as McCoy et al. (2018, 2020) tackle the poverty of the stimulus by testing the acquisition of specific language abilities (i.e., auxiliary inversion). However, the setup in those studies involves no pre-training or Language Modeling phase, therefore treating the phenomenon as a free-standing task. It is difficult to analyze reported results with respect to children acquisition theories, since, as the authors note themselves, humans tend to share processing strategies across phenomena. As mentioned above, Huebner et al. (2021) propose instead an attractive framework tested on TLMs, which is however affected by the exact hyperparameter setting of the model.

Turning to the actual shape of the input, Yu et al. (2020) investigate the grammatical judgments of NLMs in a minimal pair setting (i.e., two sentences that differ in their acceptability due to just one grammatical property). They find that performance is correlated across tasks and across models, suggesting that the *learnability* of an item does not depend on

a specific model but seems to be rather tied to the statistical properties of the input (i.e., on the distribution of constituents).

In Davis and van Schijndel (2020), the authors examine biases of ANNs for ambiguous relative clause attachments. In a sentence like *Andrew had dinner yesterday with the nephew of the teacher that was divorced*, both *nephew* and *teacher* are available for modifications by the relative clause: from a purely grammatical perspective, both interpretations are equally plausible. English speakers however have a generic preference for attaching the relative clause to the lower nominal, while other languages such as Spanish show a preference for the higher nominal. RNNs trained on either English or Spanish do not simulate this pattern, and instead consistently prefer the low attachment (similar results are reported in Davis et al. (2020) about the influence of implicit causation on syntactic representations). The authors show this preference is an artifact of training the network on production data which, in Spanish, contains more instances of low attachments. By manually correcting this bias in the input, generating an equal proportion of high and low attachments, they find that a preference for the higher nominal is learnable by the LSTM.

Lepori et al. (2020) experiment with an artificially constructed set of simple transitive sentences (Subject-Verb-Object), containing optional adjectival or prepositional modifiers in a controlled, probabilistic setting. They show that when a BiLSTM is fine-tuned on a distribution which explicitly requires moving beyond lexical co-occurrences and creating more abstract representations, performance dramatically improves: this suggests that a simple sequential mechanism can be enough if the linguistic signal is structured in a way that abstraction is encouraged.

Finally, Pannitto and Herbelot (2020) confirm the tendency of ANNs to reproduce the particular input they are exposed to. They train an LSTM on three different genres of child-directed data. Their results show that when asked to generate, the network accurately reproduces the distribution of the linguistic constituents in its training data, while showing much lower correlation with the distribution of the other two genres.

Overall, there seems to be evidence across the board that the statistical properties of the language input affect learnability as a whole and are responsible for inter-speaker differences. This fits well in a usage-based framework, and it also contributes to a view of grammar that allows for partial competence, as we will now discuss.

5. GRADED VS. DISCRETE NOTION OF GRAMMAR

Usage-based theories take a graded view on acquisition of linguistic structures, acknowledging that partial competence can be observed, blurring the distinction between semantic and syntactic knowledge, and ultimately, allowing for a range of varied grammatical intuitions across speakers. Existing studies on the grammatical abilities of RNNs report results which tend to confirm this view, but they are interpreted in different ways, as we will presently see.

Wilcox et al. (2018) address the phenomenon of filler-gap dependencies (e.g., the dependency existing between *what* and its gap in *I know what/* that the lion devoured - at sunrise*), evaluating the surprisal values assigned by the pre-trained language models of Gulordava et al. (2018) and Chelba et al. (2013). Their results show that neural language models show high peaks of surprisal in the post-gap position, irrespective of the syntactic position where the gap happens (either subject, object or prepositional phrase). When considering the whole clause, however, predictions related to the subject position are much stronger than for the other two positions, correlating with human online processing results. Overall, their results indicate that filler-gap dependencies, and the constraints on them, are acquired by language models, albeit in a graded manner, and in many cases correlate with human judgements. Similar results are reported by Chowdhury and Zamparelli (2018), but the authors commit to a stronger binary distinction between competence and performance, ultimately stating that their model "is sensitive to linguistic processing factors and probably ultimately unable to induce a more abstract notion of grammaticality."

A call for *full abstraction*, as opposed to a graded view of syntactic abilities, is also expressed in Marvin and Linzen (2018): English artificial sentence pairs (i.e., a grammatical sentence with its ungrammatical counterpart) are automatically built using a non recursive context free grammar, with the intent of minimizing "the semantic or collocational cues that can be used to identify the grammatical sentence." Two models are evaluated: a simple RNN language model and a multi-task RNN that solves two tasks at the same time, language modeling and a tagging task that superimposes syntactic information, both trained on a Wikipedia subset. Overall, results are varied both between tasks and, for a single benchmark, between different lexical items: a result that, as the authors say "would not be expected if its syntactic representations were fully abstract." The outcome is however perfectly reasonable in a usage-based framework, if we think of abstraction as induced by the association of specific lexical items with grammatical structure and intentions.

Gradedness is instead the explicit focus of Hawkins et al. (2020), where the authors examine the performance of various pre-trained neural language models, including the LSTM of Gulordava et al. (2018), against a dataset containing human preference judgements on dative alternations in various conditions, manipulating the length and definiteness of the recipient argument. In this study aimed at modeling verb biases, human intuitions are collected and kept as graded values, which the models are tested against. Lexical bias is seen here as a proxy of syntactic abilities rather than as something that might hurt the abstraction process.

Summarizing, we see a growing body of evidence for gradedness of linguistic judgements, both in humans and networks. Interestingly, studies such as Liu et al. (2021) also show that the acquisition of different types of linguistic knowledge proceeds in parallel, but at various rates, in both LSTMs and TLMs. This opens the door for thinking of the potential aggregation of syntactic and semantic knowledge, but also for talking of different levels of competence, as acquisition takes place over time.

6. DISCUSSION

The current tendency in the computational community is to give an account of the knowledge acquired at the end of the acquisition process (Linzen et al., 2018, 2019; Alishahi et al., 2019; Baroni, 2020), but the picture emerging from the analysis of NLMs linguistic abilities is variegated, both in terms of approaches and results. To some extent, the inconsistent results reported in the literature are due to differences in theoretical assumptions made by each of the mentioned studies, rather than in experimental designs. As already highlighted by Linzen and Baroni (2021), the conclusions drawn by ANNs studies largely depend on the particular notions of competence, performance, lexicon and grammar that researchers commit to. Perhaps surprisingly, very few studies explicitly link the performance of neural language models to usage-based formalisms.

More specifically, the evaluation of NLMs is widely performed over specialized datasets that capture some highly debated phenomena, such as auxiliary inversion or agreement in increasingly puzzling contexts. Datasets comprehending a wider range of phenomena are now emerging (Hu et al., 2020; Warstadt et al., 2020a). The mastery of such phenomena undoubtedly corresponds to important milestones in acquisition, but they only give a partial view on the learner's trajectory towards full productivity and compositionality. More careful investigations are required to show how biases in the input affect learning and grammatical performance, and how such biases are eventually overcome.

Another issue is that the performance of NLMs is often compared to those of adult speakers. But some usage-based theories rely on the idea that grammar is an ability that evolves throughout the human lifespan, generating different learning patterns in children and adults. To fully explore this idea, studies should increase their focus on alternative datasets, both at input and evaluation stage.

Finally, NLMs are usually treated as an idealized *average* speaker, with their predictions being compared to aggregates of human judgements. While this can be regarded as a necessary simplification, it also mirrors the view that there is a universally shared grammar towards which both speakers and LMs converge, and that this convergence, rather than individual differences, is meaningful. Conceptualizing NLMs as individual speakers rather than communities would probably let different evaluation setups emerge and provide new modeling possibilities for usage-based accounts.

AUTHOR CONTRIBUTIONS

LP prepared the literature review. AH supervised the work. LP and AH jointly wrote the survey. Both authors contributed to the article and approved the submitted version.

REFERENCES

Alishahi, A., Chrupała, G., and Linzen, T. (2019). Analyzing and interpreting neural networks for NLP: a report on the first BlackboxNLP workshop. *Nat. Lang. Eng.* 25, 543–557. doi: 10.1017/S135132491900024X

Arehalli, S., and Linzen, T. (2020). "Neural language models capture some, but not all, agreement attraction effects," in *CogSci 2020*.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375, 1. doi: 10.1098/rstb.2019.0307

Barsalou, L. W. (1987). "The instability of graded structure: implications for the nature of concepts," in *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, ed U. Neisser, Barsalou 1983, New York, NY: Cambridge University Press, 101–140.

Boyd, J. K., and Goldberg, A. E. (2009). Input effects within a constructionist framework. *Mod. Lang. J.* 93, 418–429. doi: 10.1111/j.1540-4781.2009.00899.x

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., et al. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv [cs.CL]*.

Chowdhury, S. A., and Zamparelli, R. (2018). "RNN simulations of grammaticality judgments on long-distance dependencies," in *Proceedings of the 27th International Conference on Computational Linguistics* (Association for Computational Linguistics) Santa Fe, NM, 133–144.

Christiansen, M. H. (2019). Implicit statistical learning: a tale of two literatures. *Top. Cogn. Sci.*, 11, 468–481. doi: 10.1111/tops.12332

Christiansen, M. H., and Chater, N. (2015). The now-or-never bottleneck: a fundamental constraint on language. *Behav. Brain Sci.* 39, 1–72. doi: 10.1017/S0140525X1500031X

Christiansen, M. H., and Chater, N. (2016). *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

Cornish, H., Dale, R., Kirby, S., and Christiansen, M. H. (2017). Sequence memory constraints give rise to language-like structure through iterated learning. *PLoS ONE* 12, 1–18. doi: 10.1371/journal.pone.0168532

Davis, F., and van Schijndel, M. (2020). Discourse structure interacts with reference but not syntax in neural language models. *Proc. 24th Conf. Comput. Nat. Lang. Learn.* 396–407. doi: 10.18653/v1/2020.conll-1.32

Davis, F., and van Schijndel, M. (2020). "Recurrent neural network language models always learn English-like relative clause attachment," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1979–1990.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). "Recurrent neural network grammars," in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference* (Association for Computational Linguistics (ACL)), 199–209.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cogn. Sci.* 33, 547–582. doi: 10.1111/j.1551-6709.2009.01023.x

Fazekas, J., Jessop, A., Pine, J., and Rowland, C. (2020). Do children learn from their prediction mistakes? a registered report evaluating error-based theories of language acquisition. *R. Soc. Open Sci.* 7, 180877. doi: 10.1098/rsos.180877

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., and Zuidema, W. (2018). "Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Brussels), 240–248.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. New York, NY: Oxford University Press.

Gomez, R. L., and Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70, 109–135.

Gómez, R. L., and Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends Cogn. Sci.* 4, 178–186. doi: 10.1016/S1364-6613(00)01467-4

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). "Colorless green recurrent networks dream hierarchically," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 1195–1205.

Hart, B., Risley, T. R., and Kirby, J. R. (1997). Meaningful differences in the everyday experience of young american children. *Can. J. History Sport Phys. Educ.* 22, 323.

Hawkins, R. D., Yamakoshi, T., Griffiths, T. L., and Goldberg, A. E. (2020). "Investigating representations of verb bias in neural language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics), 4653–4663.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). "A systematic assessment of syntactic generalization in neural language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics), 1725–1744.

Huebner, P. A., Sulem, E., Cynthia, F., and Roth, D. (2021). "BabyBERTa: Learning more grammar with small-scale child-directed language," in *Proceedings of the 25th Conference on Computational Natural Language Learning* (Punta Cana: Association for Computational Linguistics), 624–646.

Jackendoff, R. (2002). *Foundations of Language*. New York, NY: Oxford University Press.

Kharitonov, E., Baroni, M., and Hupkes, D. (2021). How bpe affects memorization in transformers. *arXiv preprint* arXiv:2110.02782.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., and Smith, N. A. (2017). "What do recurrent neural network grammars learn about syntax?" in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1 (Valencia: Association for Computational Linguistics), 1249–1258.

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., and Blunsom, P. (2018). "LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1 (Melbourne, NSW: Association for Computational Linguistics), 1426–1436.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. (2019). "The emergence of number and syntax units in LSTM language models," in *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)*, vol. 1 (Minneapolis, MN: Association for Computational Linguistics), 11–20.

Lepori, M. A., Linzen, T., and McCoy, T. R. (2020). "Representations of syntax mask useful: Effects of constituency and dependency structure in recursive lstms," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics), 3306–3316.

Linzen, T., and Baroni, M. (2021). Syntactic structure from deep learning. *Ann. Rev. Linguist.* 7, 1–19. doi: 10.1146/annurev-linguistics-032020-051035

Linzen, T., Chrupala, G., and Alishahi, A. (2018). *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels: Association for Computational Linguistics. Available online at: https://aclanthology.org/W18-5400

Linzen, T., Chrupala, G., Belinlov, Y., and Hupkes, D. (2019). *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence: Association for Computational Linguistics. Available online at: https://aclanthology.org/W19-4800

Liu, Z., Wang, Y., Kasai, J., Hajishirzi, H., and Smith, N. A. (2021). "Probing across time: what does roberta know and when?" in *Findings of the Association for Computational Linguistics: EMNLP 2021* (Punta Cana), 820–842.

Marvin, R., and Linzen, T. (2018). "Targeted syntactic evaluation of language models," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels: Association for Computational Linguistics), 1192–1202.

McCoy, R. T., Frank, R., and Linzen, T. (2018). "Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks," in *CogSci*, eds T. Rogers, (Madison, WI: The Cognitive Science Society), 2096–2101.

McCoy, R. T., Frank, R., and Linzen, T. (2020). Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Trans. Assoc. Comput. Linguist.* 8, 125–140. doi: 10.1162/tacl_a_00304

McRae, K., and Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Lang. Linguist. Compass* 3, 1417–1429. doi: 10.1111/j.1749-818X.2009.00174.x. People

Pannitto, L., and Herbelot, A. (2020). "Recurrent babbling: evaluating the acquisition of grammar from limited input data," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, 165–176.

Pickering, M. J., and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347. doi: 10.1017/S0140525X12001495

Ramscar, M., Dye, M., and McCauley, S. M. (2013). Error and expectation in language learning: the curious absence of mouses in adult speech. *Language* 89, 760–793. doi: 10.1353/lan.2013.0068

Romberg, A. R., and Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdiscipl. Rev. Cogn. Sci.* 1, 906–914. doi: 10.1515/9781934078242

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Vaswani, A., Shazeer, N., Parmer, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Curran Associates). Available online at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., et al. (2020a). Blimp: the benchmark of linguistic minimal pairs for english. *Trans. Assoc. Comput. Linguist.* 8, 377–392. doi: 10.1162/tacl_a_00321

Warstadt, A., Zhang, Y., Li, X., Liu, H., and Bowman, S. R. (2020b). "Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually)," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: Association for Computational Linguistics).

Warstadt, A., and Bowman, S. R. (2020). "Can neural networks acquire a structural bias from raw linguistic data?" in *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020*, eds S. Denison, M. Mack, Y. Xu, and B. C. Armstrong, cognitivesciencesociety.org.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). "What do RNN language models learn about filler gap dependencies?" in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. (Association for Computational Linguistics).

Yu, C., Sie, R., Tedeschi, N., and Bergen, L. (2020). "Word frequency does not predict grammatical knowledge in language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics), 4040–4054.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.