PERSPECTIVES

# Fusion is great, and interpretable fusion could be exciting for theory generation: Response to Pater

LISA S. PEARL

*University of California, Irvine*

From my perspective, Pater's (2019) target article does a great service both to researchers who work in generative linguistics and to researchers who utilize neural networks—and especially to researchers who might find themselves wanting to do both by harnessing the insights of each tradition. The fusion of theories of linguistic representation and probabilistic learning techniques has certainly led to many interesting and valuable insights about the nature of both linguistic representation and the language acquisition process. However, I feel that the most exciting aspect of Pater's article is the increasing interpretability of neural network models, especially when combined with insights from the theoretical framework of generative linguistics. This allows for the possibility that neural networks could be used to actually generate new theories of representation. I describe how I think this theory-generation process might work with interpretable neural networks.

*Keywords*: generative linguistics, interpretable neural networks, neural networks, probabilistic learning models, symbolic models, theory evaluation, theory generation

**1.** FUSION IS GREAT FOR THEORY. Joe Pater's (2019) article highlights that learning theories are intimately related to theories of linguistic representation. That is, one reason why we theorize that representations take certain forms is because of how these representations are learned by children (and importantly, what the constraints are on that learning process). This learnability-centric approach complements an approach based on typological economy, whose goal is to identify a representation that accounts for all and only the variation observed in human languages. In my view, the typological-economy approach is a fine way to generate theories of possible representations, and the learnability approach is a fine way to evaluate theories of possible representations. Pater notes that this is a productive cycle in phonology, where OPTIMALITY THEORY and the HARMONIC GRAMMAR theory of typology are supported relative to the success of the learning algorithms associated with them; I have also taken this learnability approach in some of my recent work comparing parametric theories of phonology (Pearl 2017, Pearl, Ho, & Detrano 2017). More generally, I think there is great value in making 'arguments from acquisition' to evaluate proposed theoretical representations, which I have done in other recent work (e.g. Pearl & Sprouse 2013, 2018a,b).

This link between linguistic representation and language acquisition means we need to seriously consider what it is that human learning mechanisms are capable of, given the data children encounter. Only then can we evaluate possible linguistic representations that encode the linguistic knowledge children attain after learning. As Pater notes, this is especially relevant for considerations of 'how much and what kind of explicitly prespecified linguistic structure is needed' (p. e43), which is the heart of the UNIVERSAL GRAMMAR (UG) debate. There are two key findings from recent symbolic probabilistic modeling work that speak to exactly this (Perfors, Tenenbaum, & Regier 2011, Pearl & Sprouse 2013, 2018a, Pearl & Mis 2016, Abend et al. 2017): (i) children can get by without the built-in language-specific knowledge that we previously thought they needed, but (ii) children still need something built in (which is often an ability to generate structured representations of particular kinds).

The findings I mentioned above are primarily in the realm of syntax, but Pater notes that certain representational analyses in phonology were previously excluded for learnability reasons. Now that we have a better understanding of the learning mechanisms children can use, these analyses are probably worth reconsidering. From the perspective of what is in UG, probabilistic modeling allows us to investigate the explicitly prespecified linguistic structure that is needed for specific representations to be learnable by children.

Where I initially diverged from Pater's view on the value of neural networks was exactly what to use them FOR: that is, what added value did they have over other probabilistic learning techniques? As I noted above, my own work and that of many others in the language acquisition modeling community relies on probabilistic inference over symbolic representations. This contrasts with the distributed representations that neural networks typically rely on, which are often far more difficult to interpret.

Maybe neural networks were meant as an implementational-level model of a representation and learning process (in the sense of Marr 1982), mimicking how the wetware of the brain would implement the symbolic representations and inference of prior models. If so, one clear added value of implementational-level neural network models is demonstrating how the proposed symbolic representation and learning process can be carried out in human brains.

Yet, from what I could tell, this did not seem to be what the neural networks used in language modeling were doing. Many of them (especially those used in deep-learning approaches) do not share a lot of implementational details with the human brain as we now understand brains, in contrast with neuromorphic models that encapsulate neurobiology in great detail (e.g. Neftci et al. 2013, Krichmar, Conradt, & Asada 2015, Avery & Krichmar 2017, Beyeler et al. 2017, Neftci et al. 2017). So, the neural networks of language modeling are another instance of a higher-level probabilistic learning model—and typically an idealized, computational-level model at that, if the neural networks do not take the learning-period limitations of human language acquisition into account. This means that these neural network models are not telling us how human brains implement linguistic representation and learning; instead, they are telling us what representations and learning procedures could give rise to observable linguistic behavior we care about, given certain assumptions about probabilistic learning. That is, these neural network models would be used for theory evaluation, the same as traditional symbolic probabilistic models.

But this again highlights a key downside to neural networks as we knew them: they are really hard to interpret. If I am a cognitive modeler who wants to evaluate a particular theory that is typically stated in symbolic terms, why should I choose a neural network whose innards are hard to interpret over some other probabilistic learning technique (e.g. Bayesian inference) whose innards are easy to interpret?

The answer came to me near the end of Pater's article: I don't. That is, I do not want to use neural networks for explicit theory evaluation—if I already have an explicit theory, I am better off with a model that is easy to interpret. However, what if I don't already have an explicit theory? What if instead I have a sense that there are certain necessary building blocks, but I do not quite know if the ones I am thinking of can be stuck together by a plausible probabilistic learner in the right way to yield human language acquisition behavior?

## 2. Interpretable fusion could be exciting for theory generation.

Pater characterizes recurrent neural networks (RNNs) the following way: RNNs are 'given the structural building blocks of symbols and their roles, but must learn their configurations' (p. e63). To me, this bears a striking resemblance to explicit hypothesis construction from a latent hypothesis space defined by certain building blocks and their roles (Perfors 2012)—and that is something that generative linguists are already familiar with. For instance, linguistic parameters are the building blocks that define a latent hypothesis space of grammars (e.g. $n$ binary parameters define a latent hypothesis space of $2^n$ explicit grammars); a probabilistic learning algorithm like variational learning (Yang 2002, 2004, 2012) is a way to generate explicit grammar hypotheses and evaluate them, given the data children encounter. Another recent example from my own work is using a cognitively motivated decision criterion called the TOLERANCE PRINCIPLE (Yang 2005, 2016) to construct explicit linking-theory hypotheses from a latent hypothesis space defined over links between syntactic positions and thematic roles (Pearl & Sprouse 2018b). These learning mechanisms have in common that they can generate an explicit hypothesis from the predefined building blocks in the latent hypothesis space—but the explicit hypothesis ITSELF is not explicitly predefined.

I think the same intuition holds for overhypotheses in hierarchical Bayesian inference (see Pearl & Goldwater 2016 and Pearl 2019 for an accessible overview). The overhypothesis defines the building blocks out of which the more specific hypotheses are constructed. As a quick nonlinguistic example of overhypotheses (derived from an experiment with nine-month-olds by Dewar and Xu (2010)), consider an infant who is trying to figure out the contents of bags in an experimental setup. From the first bag, the experimenter pulls out four circles of different colors; from the second bag, four squares of different colors; and from the third bag, four triangles of different colors. Now, from the fourth bag, the experimenter pulls out a single star. What should the rest of the fourth bag contain? Adults (and nine-month-olds) expect it to contain three more stars of different colors. Why? The infant has never seen stars before in this experimental setup—why should she have any expectations at all about bags with stars? The answer is that she has formed an overhypothesis about the contents of bags, on the basis of the first three example bags: bags contain four objects of the same SHAPE (all of different colors). So, when she sees a star come out of the fourth bag, she can generate the explicit hypothesis that the fourth bag contains four STARS all of different colors. If an experimenter reached into a fifth bag and pulled out a pentagon, the infant would generate the explicit hypothesis that the fifth bag contains four PENTAGONS all of different colors. This is the same generative power we saw before: the ability to generate an explicit hypothesis to be evaluated, on the basis of available building blocks. In this case, the building blocks are supplied by the overhypothesis, with ‘shape’, ‘4’, and ‘all of different colors’ as building blocks in that overhypothesis.

In fact, hierarchical Bayesian inference has the ability to have overhypotheses on overhypotheses (called ‘over-over-hypotheses’), so that the building blocks of one level may be the explicit hypotheses from a previous level. In the example above about the contents of bags, the over-over-hypothesis might contain the building block of ‘exact number’, which would allow generation of the overhypothesis of ‘4’ (but also ‘1’ or ‘2’ or ‘7’); similarly, it might contain the building block of ‘color distribution’, which would allow generation of the overhypothesis of ‘all of different colors’ (but also ‘all the same color’ or ‘50–50 color split’ or ‘90–5–5 color split’). In this way, hierarchical Bayesian inference could allow us to identify very basic building blocks that could be configured into more and more specific explicit hypotheses.

However, a current limitation is that we, as modelers, have to specify both those core building blocks and, for every level of overhypothesis we have, how an explicit hypothesis at the next level is generated. In practical terms, this means we are constrained by current symbolic mathematical methods that define how this explicit hypothesis-generation process occurs. This also means that we are constrained by our own conceptions of what the building blocks could be and, importantly, how explicit hypotheses could be generated from those building blocks. Let us consider the example about the contents of bags again: what are the sufficient building blocks of the over-over-over-hypothesis that allow generation of ‘exact number’ and ‘color distribution’? Will certain conceptual features do? What about attentional features?

For me, this is where RNNs come in. Say we have some ideas about potential building blocks at a very fundamental level (e.g. the equivalent of the over-over-over-hypothesis). However, we do not know if these building blocks are capable of generating reasonable over-over-hypotheses, that in turn generate reasonable overhypotheses, that in turn generate reasonable hypotheses, that in turn lead to observable linguistic behavior. This could be where an RNN could shine. The RNN would take in those fundamental building blocks and identify configurations that could lead to observable behavior WITHOUT the modeler having to explicitly tell the model how to generate the intermediate hypothesis levels (though of course the modeler has to specify the RNN architecture).

Why is this any better than using hierarchical Bayesian modeling? I think it is because an RNN might generate intermediate hypotheses (e.g. over-over-hypotheses and overhypotheses) that are not something a human modeler would naturally think of, given those original building blocks. The RNN generates intermediate hypotheses based on whatever architectural biases mold its explicit hypothesis generation; so, it may give higher probability to explicit hypotheses that were very low probability for a hierarchical Bayesian model. Yet these novel explicit hypotheses might well be a viable option for generating observable linguistic behavior—they are just not ones a human modeler would think of.

I think a helpful analogy comes from the related fields of genetic/evolutionary algorithms and evolutionary programming (Koza 1994, 1995, Back 1996, Li et al. 2015, Khan 2018); both have been used (among other ways) for optimization problems, such as designing circuits. These algorithms ‘evolve’ a solution to a problem using biologically inspired manipulations of the relevant building blocks, and they not only can find the best solutions, but also can find ones that humans did not think of. That is, because human problem solvers (e.g. circuit designers) are implicitly fettered by ideas about what the solution will look like, they end up exploring only a part of the latent hypothesis space. This, of course, is a part of the hypothesis space that, from their perspective, has a high probability of containing the best solution. But that is the point: it is FROM THEIR PERSPECTIVE, given whatever biases they have, based on their knowledge of what existing solutions look like and the principles that allowed them to generate those solutions.

I think the same could be true for theories of linguistic representation. We could be implicitly fettered by our ideas about what the representation will look like, given certain building blocks, existing theories, and how we generated previous theories of representation from those building blocks. So because of that, we explore only certain parts of the latent hypothesis space of representations that seem like high-probability regions to us—and we do this by symbolically encoding how the building blocks from one level generate the explicit hypotheses of the next level. In contrast, an RNN could be given the same fundamental building blocks and then explore a different section of the latent hypothesis space, implicitly fettered by its architecture. Importantly, its architecturally based fetters are not obviously the same as our (symbolic) idea-based fetters, and so perhaps an RNN can identify representation possibilities we have not yet conceived of that nonetheless match observable linguistic behavior.

If that happens, then everything hinges on how interpretable the RNN solutions are. If we can interpret a novel solution—that is, translate it back into symbolic terms we can understand—we then have a new proposal for a theory of linguistic representation, based on the building blocks we gave the RNN. As Pater notes, there are equivalencies between the distributed representations of existing RNN models and linguistic structures that linguists recognize (e.g. Palangi et al. 2017). This gives me hope that the distributed representation solutions that future RNNs come up with may lead to possible linguistic representations we have not considered before—and this is exciting indeed.

## REFERENCES

Abend, Omri; Tom Kwiatkowski; Nathaniel J. Smith; Sharon Goldwater; and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition* 164.116–43. DOI: 10.1016/j.cognition.2017.02.009.

Avery, Michael C., and Jeffrey L. Krichmar. 2017. Neuromodulatory systems and their interactions: A review of models, theories, and experiments. *Frontiers in Neural Circuits* 11:108. DOI: 10.3389/fncir.2017.00108.

Back, Thomas. 1996. *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. Oxford: Oxford University Press.

Beyeler, Michael; Emily Rounds; Kristofor Carlson; Nikil Dutt; and Jeffrey L. Krichmar. 2017. Sparse coding and dimensionality reduction in cortex. *bioRxiv* 149880 (preprint). DOI: 10.1101/149880.

Dewar, Kathryn M., and Fei Xu. 2010. Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science* 21.1871–77. DOI: 10.1177/0956797610388810.

Khan, Gul Muhammad. 2018. Evolutionary computation. *Evolution of artificial neural development: In search of learning genes*, 29–37. Dordrecht: Springer. DOI: 10.1007/978-3-319-67466-7_3.

Koza, John R. 1994. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing* 4.87–112. DOI: 10.1007/BF00175355.

Koza, John R. 1995. Survey of genetic algorithms and genetic programming. *Proceedings of WESCON'95*, 589–94. DOI: 10.1109/WESCON.1995.485447.

Krichmar, Jeffrey L.; Jörg Conradt; and Minoru Asada. 2015. Neurobiologically inspired robotics: Enhanced autonomy through neuromorphic cognition. *Neural Networks* 72.1–2. DOI: 10.1016/j.neunet.2015.11.004.

Li, Bingdong; Jinlong Li; Ke Tang; and Xin Yao. 2015. Many-objective evolutionary algorithms: A survey. *ACM Computing Surveys (CSUR)* 48(1):13. DOI: 10.1145/2792984.

Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Neftci, Emre O.; Charles Augustine; Somnath Paul; and Georgios Detorakis. 2017. Event-driven random back-propagation: Enabling neuromorphic deep learning machines. *Frontiers in Neuroscience* 11:324. DOI: 10.3389/fnins.2017.00324.

Neftci, Emre O.; Jonathan Binas; Ueli Rutishauser; Elisabetta Chicca; Giacomo Indiveri; and Rodney J. Douglas. 2013. Synthesizing cognition in neuromorphic electronic systems. *Proceedings of the National Academy of Sciences of the United States of America* 110(37).E3468–E3476. DOI: 10.1073/pnas.1212083110.

Palangi, Hamid; Paul Smolensky; Xiaodong He; and Li Deng. 2017. Question-answering with grammatically-interpretable representations. arXiv:1705.08432 [cs.CL]. Online: https://arxiv.org/abs/1705.08432.

Pater, Joe. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language* 95(1).e41–e74.

Pearl, Lisa. 2017. Evaluation, use, and refinement of knowledge representations through acquisition modeling. *Language Acquisition* 24(2).126–47. DOI: 10.1080/10489223.2016.1192633.

Pearl, Lisa. 2019. Modeling syntactic acquisition. *The Oxford handbook of experimental syntax*, ed. by Jon Sprouse. Oxford: Oxford University Press, to appear.

Pearl, Lisa, and Sharon Goldwater. 2016. Statistical learning, inductive bias, and Bayesian inference in language acquisition. *The Oxford handbook of developmental linguistics*, ed. by Jeffrey Lidz, William Snyder, and Joe Pater, 664–95. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199601264.013.28.

Pearl, Lisa; Timothy Ho; and Zephyr Detrano. 2017. An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech. *Language Acquisition* 24(4).307–42. DOI: 10.1080/10489223.2016.1194422.

Pearl, Lisa, and Benjamin Mis. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric *one*. *Language* 92(1).1–30. DOI: 10.1353/lan.2016.0006.

Pearl, Lisa, and Jon Sprouse. 2013. Computational models of acquisition for islands. *Experimental syntax and island effects*, ed. by Jon Sprouse and Norbert Hornstein, 109–31. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139035309.006.

Pearl, Lisa, and Jon Sprouse. 2018a. Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. Irvine: University of California, Irvine, ms. Online: https://ling.auf.net/lingbuzz/003913.

Pearl, Lisa, and Jon Sprouse. 2018b. The acquisition of linking theories: A tolerance principle approach to learning UTAH and rUTAH. Irvine: University of California, Irvine, ms. Online: https://ling.auf.net/lingbuzz/004088.

Perfors, Amy. 2012. Bayesian models of cognition: What's built in after all? *Philosophy Compass* 7(2).127–38. DOI: 10.1111/j.1747-9991.2011.00467.x.

Perfors, Amy; Joshua B. Tenenbaum; and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118.306–38. DOI: 10.1016/j.cognition.2010.11.001.

Yang, Charles D. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, Charles D. 2004. Universal grammar, statistics or both? *Trends in Cognitive Science* 8(10).451–56. DOI: 10.1016/j.tics.2004.08.006.

Yang, Charles D. 2005. On productivity. *Linguistic Variation Yearbook* 5.265–302. DOI: 10.1075/livy.5.09yan.

Yang, Charles D. 2012. Computational models of syntactic acquisition. *WIREs Cognitive Science* 3.205–13. DOI: 10.1002/wcs.1154.

Yang, Charles D. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge, MA: MIT Press.

[lpearl@uci.edu]

[Received 30 July 2018;
revision invited 26 August 2018;
revision received 7 September 2018;
accepted 29 September 2018]