Defragmenting learning

Vsevolod Kapatsinski

Department of Linguistics
University of Oregon

Keywords: learning theory, language acquisition, cue competition, associative interference, Optimality Theory, Rescorla-Wagner model

ABSTRACT: In the 1990s, language acquisition researchers and theoretical linguists developed an interest in learning mechanisms, and learning theorists rediscovered the verbal learning tradition. Nonetheless, learning theory and language acquisition continued to develop largely independently, which has stymied progress in both fields. However, exciting progress is happening in applying learning theory to language, and, more recently, in using language learning data to advance domain-general learning theory. These developments raise hopes for a bidirectional flow of information between the fields. The importance of language data for learning theory and of learning theory for understanding language is briefly discussed.

In 1996, Elman and Bates published a *Science* editorial called *Learning rediscovered* heralding Saffran et al.'s (1996) discovery that infants show surprise when they hear a syllable that's relatively unexpected, given the preceding context. The paper has garnered thousands of citations and established a whole new field in language acqusition. Yet, as Elman and Bates pointed out, the central finding was quite unsurprising in the context of domain-general learning theory (see also Baayen et al., 2013). The fact that it was surprising enough to students of language acquisition to spark a revolution is a sign of how fragmented the study of learning had become.

In theoretical linguistics, learning was also being rediscovered since the mid-1980s, prompted to a large extent by the re-emergence of connectionism (Rumelhart et al., 1986). On the functionalist side of the field, usage-based linguistics made learning (and especially the, still-understudied, learning by doing) central to linguistic theory (Bybee, 1985; Langacker, 1987; see also Bybee & McClelland, 2005). The usage-based turn in functionalist linguistics led to an explosion of interest in frequency effects, with frequency understood as estimating the number of episodes of a particular type of experience. However, until recently, usage-based linguists have not produced explicit models of learning (McCauley & Christiansen, 2019).

On the generative side of linguistics, interest in learning exploded in the 1990s with the emergence of Optimality Theory (Prince & Smolensky, 1993 / 2004). Optimality Theory (OT) combined a connectionist understanding of grammar as a constraint satisfaction system with the strong nativism characteristic of generative grammar, by assuming that the constraints are innate. Ironically, OT's commitment to nativism stimulated an interest in learning; perhaps, by making it seem that the problem of language acquisition is on the cusp of being finally solved. This resulted in a large body of computationally explicit work on learning mechanisms.

The history of research on learning within OT is instructive. Learning in this framework began from a well-founded domain-general approach to learning, error-driven constraint reweighting (e.g., Smolensky & Legendre, 1990), later shown to be equivalent to multiple regression (e.g., Zymet, 2019). However, it was then abandoned in favor of a domain-specific strict ranking approach. This decision was based on how languages were assumed to work (Pater, 2009), and became widely popular based on the relative ease with which constraints can be ranked, rather than weighted, by hand. Subsequent work has shown the strict ranking assumption to be empirically incorrect (e.g., Sorace & Keller, 2005), which led to re-emergence of algorithms based on constraint weighting (Pater, 2009). More recently, research in this framework has come back full circle to converge with multiple regression (Hayes & Wilson, 2008; Zymet, 2019). In the end, although much important work has been done, thirty years of research have brought the field back to where it started.

What went wrong? I would argue that the root of the problem is fragmentation, caused by an assumption that language learning is so fundamentally different from other kinds of learning that it demands fundamentally different inference algorithms (see Chomsky, 1965, for an explicit argument to this effect), and that therefore we could safely ignore the huge body of work on how we learn in other domains and start anew.

As Miller and Escobar (2002) noted, the assumption that language learning is special has also done damage to the study of learning outside of language. Thus, the rather large literature on associative interference in verbal learning was entirely ignored in the subsequent development of animal learning models, which has led these models to assume, as a central claim, that cues must co-occur in order to compete (e.g., Rescorla & Wagner, 1972). This assumption, which the verbal learning literature contradicted, was then shown not to hold in both causal learning in humans (Matute & Pineño, 1998a) and, more impressively, classical conditioning in animals (Matute & Pineño, 1998b). More generally, the decades-long isolation of learning theory from language acquisition has meant not only that the insights of learning theory are often ignored in language acquisition, but also that learning theory has come to depend on a relatively small number of experimental paradigms, especially in humans (see also Chartier & Fagot, 2022). And even within learning theory, fragmentation has made it difficult to build on prior work. For example, Powell et al. (2016), working on causal learning in a Bayesian paradigm, propose competition between non-co-occurring cues as a novel, untested prediction.

In the last decade, psycholinguists have started to use domain-general learning theory to model language acquisition. The models being applied go beyond regression, which describes how cues ought to be weighted at the end of learning, and predict what will be learned from an individual trial, or even an individual second of experience (Nixon & Tomaschek, 2021). In particular, the Rescorla-Wagner model (1972) has provided elegant accounts of many puzzling phenomena in language acquisition and language processing (e.g., Arnon & Ramscar, 2012; Baayen et al., 2011; Ellis, 2006; Kapatsinski, 2018; Ramscar et al., 2013b, 2014), and generated several accurate novel predictions (e.g., Nixon, 2020; Olejarczuk et al., 2018; Ramscar et al., 2010, 2013a). Nonetheless, until recently, psycholinguists have not compared alternative learning-theoretic models. This work has therefore succeeded in establishing that learning theory can provide insights into language acquisition but has not shown that language acquisition can also inform learning theory. For example, it is only recently that the *unique* predictions that distinguish Rescorla-Wagner from other theories of learning have begun to be investigated in language learning (Caballero & Kapatsinski, 2022; Harmon et al., 2019; Kapatsinski, 2021; Nixon et al., 2022). I hope that this trend will continue, and the next decade will see a *bidirectional* flow of information between language learning and domain-general learning theory.

Why is language acquisition an important domain for learning theory to explain? First, the statistics of the linguistic environment are relatively well understood because of the availability of large corpora, and these statistics have some distinctive characteristics. For example, there is a usually a very large number of possible outcomes in any context, although rich-get-richer loops in language change mean that there is a small number of knock-out cues often have categorical effects; cue co-occurrence structures are often degenerate; frequency distributions are highly skewed; and superadditive interactions between cues are common (Breiss & Albright, 2022; Kapatsinski, 2013; Zipf, 1949). These characteristics are quite different from the statistics in typical learning experiments (see also Wasserman et al., 2015) but may be shared with the statistics of experience in other domains. For example, competition between non-co-occurring cues means that form-meaning connectivity and cause-and-effect relationships (Luque et al., 2017; Powell et al., 2016) are likely to be characterized by sparse-and-strong structure.

Theory development must, however, be guided both by well-documented characteristics of the object of learning, and what we already know about how learning works. That is, to use a result to develop a new theory of learning, the researcher should need to demonstrate that existing plausible theories cannot account for the results. Second, language is a complex suite of behaviors that requires the whole brain (as well as the body), and the brain implements more than one learning mechanism (McClelland et al., 1995). Although some theories are alternative accounts of the same mechanism, others describe distinct coexisting mechanisms. For example, Hebbian and error-driven learning are often treated as alternatives in psycholinguistics but likely coexist in the brain (Ashby et al., 1998). An important direction for future work is therefore also to investigate how these multiple mechanisms work together in acquiring language and other complex skills.

References:

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292-305.

Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442-481.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech*, *56*(3), 329-347.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438-481.

Bates, E., & Elman, J. (1996). Learning rediscovered. *Science*, *274*(5294), 1849-1850.

Breiss, C., & Albright, A. (2022). Cumulative markedness effects and (non-)linearity in phonotactics. *Glossa*, *7*(1).

Bybee, J. (1985). *Morphology*. John Benjamins.

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, 22(2-4), 381-410.

Caballero, G., & Kapatsinski, V. (2022). How agglutinative? Searching for cues to meaning in Choguita Rarámuri (Tarahumara) using discriminative learning. In A. D. Sims, A. Ussishkin, J. Parker & S. Wray (Eds.), *Morphological diversity and linguistic cognition* (pp.121-159). Cambridge, UK: Cambridge University Press.

Chartier, T. F., & Fagot, J. (2022). Associative symmetry: A divide between humans and nonhumans? *Trends in Cognitive Sciences*.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164-194.

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76-88.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379-440.

Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language*, 89(1), 110-148.

Kapatsinski, V. (2018). *Changing minds changing tools: From learning theory to language acquisition to language change*. MIT Press.

Kapatsinski, V. (2021). Learning fast while avoiding spurious excitement and overcoming cue competition requires setting unachievable goals: Reasons for using the logistic activation function in learning to predict categorical outcomes. *Language, Cognition and Neuroscience*, 1-22.

Langacker, R. W. (1987). *Cognitive grammar.* Stanford, CA: Stanford University Press.

Luque, D., Morís, J., López, F. J., & Cobos, P. L. (2017). Previously acquired cue–outcome structural knowledge guides new learning: Evidence from the retroactive-interference-between-cues effect. *Memory & Cognition*, *45*(6), 916-931.

Matute, H., & Pineño, O. (1998a). Cue competition in the absence of compound training: Its relation to paradigms of interference between outcomes. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory,* Vol. 38, pp. 45–81). Academic Press.

Matute, H., & Pineño, O. (1998b). Stimulus competition in the absence of compound conditioning. *Animal Learning & Behavior*, *26*(1), 3-14.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1-51.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419-457.

Miller, R. R., & Escobar, M. (2002). Associative interference between cues and between outcomes presented together and presented apart: An integration. *Behavioural Processes*, *57*(2-3), 163-185.

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081.

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, *212*, 104697.

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard*, *4*(s2).

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, *33*(6), 999-1035.

Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive Psychology*, *86*, 62-86.

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Ramscar, M., Dye, M., & Klein, J. (2013a). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017-1023.

Ramscar, M., Dye, M., & McCauley, S. M. (2013b). Error and expectation in language learning: The curious absence of "mouses" in adult speech. *Language*, 89(4), 760-793.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5-42.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909-957.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (eds.), *Classical conditioning II: Current research and theory* (pp.64-99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing*. MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497-1524.

Wasserman, E. A., Brooks, D. I., & McMurray, B. (2015). Pigeons acquire multiple categories in parallel via associative learning: A parallel to human word learning? *Cognition*, *136*, 99-122.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.

Zymet, J. (2019). Learning a frequency-matching grammar together with lexical idiosyncrasy: MaxEnt versus hierarchical regression. In *Proceedings of the Annual Meetings on Phonology* (Vol. 6).