This paper will appear as a *Perspectives* article in *Language*. The *Language* editorial team is inviting responses to this paper to published along with it. For full details, see this link: https://journals.linguisticsociety.org/language/index.php/language/pater

Generative linguistics and neural networks at 60: foundation, friction, and fusion*

Joe Pater, University of Massachusetts Amherst
May 29, 2018.

Abstract. The birthdate of both generative linguistics and neural networks can be taken as 1957, the year of the publication of foundational work by both Noam Chomsky and Frank Rosenblatt. This paper traces the development of these two approaches to cognitive science, from their largely autonomous early development in their first thirty years, through their collision in the 1980s around the past tense debate (Rumelhart and McClelland 1986, Pinker and Prince 1988), and their integration in much subsequent work up to the present. Although this integration has produced a considerable body of results, the continued general gulf between these two lines of research is likely impeding progress in both: on learning in generative linguistics, and on the representation of language in neural modeling. The paper concludes with a brief argument that generative linguistics is unlikely to fulfill its promise of accounting for language learning if it continues to maintain its distance from neural and statistical approaches to learning.

1. Introduction

At the beginning of 1957, two men nearing their 29th birthdays published work that laid the foundation for two radically different approaches to cognitive science. One of these men, Noam Chomsky, continues to contribute sixty years later to the field that he founded, generative linguistics. The book he published in 1957, *Syntactic Structures*, has been ranked as *the* most influential work in cognitive science from the 20th century.[1] The other one, Frank Rosenblatt, had by the late 1960s largely moved on from his research on perceptrons – now called neural networks – and died tragically young in 1971. On the same list of 100 influential works that ranked *Syntactic Structures* #1, there is nothing by Rosenblatt, though *Perceptrons,* the 1969 book by Minsky and Papert, is listed at #74. Rosenblatt's 1957 tech report, "The perceptron: a perceiving and recognizing

* Acknowledgements: Thank you to Emily Bender, Sam Bowman, Matt Goldrick, Mark Liberman, Fred Mailhot, Joe Pater (*père*) and Language editors Andries Coetzee and Megan Crowhurst for very useful comments on an earlier version of this manuscript. For helpful discussion, thank you to Ricardo Bermúdez-Otero, Noam Chomsky, Brian Dillon, Amanda Doucette, Robert Frank, Yoav Goldberg, Thomas Graf, Gaja Jarosz, Kyle Johnson, Samuel Jay Keyser, Andrew Lamont, Tal Linzen, George Nagy, Brendan O'Connor, Barbara Partee, Brandon Prickett, Alan Prince, Paul Smolensky, Emma Strubell, Aaron Traylor and Kristine Yu, and participants in the Workshop on Perceptrons and Syntactic Structures at 60. Supported by NSF grant BCS-1650957 to the University of Massachusetts Amherst.

[1] The list "The Top 100 most influential works in cognitive science from the 20th century" was compiled by the Center for Cognitive Sciences, University of Minnesota, in 2000 (Sanz 2008). I realize that picking any date, and any person, as the beginning of a tradition is somewhat arbitrary. Many of the ideas in Chomsky (1957) and Rosenblatt (1957) can be traced back much earlier (Pullum 2011; Schmidhuber 2015). But Chomsky and Rosenblatt's work was clearly particularly prominent and influential at the time, and sowed the seeds for much subsequent research in generative linguistics and neural networks.

automaton", is very short and fairly programmatic, but the line of research that it began, much of it presented in his 1962 book, *Principles of Neurodynamics*, launched modern neural network modeling, including the multi-layer perceptrons used in deep learning (LeCun, Bengio & Hinton 2015).

Chomsky and Rosenblatt's approaches to cognitive science were radically different partly because they started at opposite ends of the problem of developing a computational theory of the human mind. Chomsky took a "high level" cognitive phenomenon – language, and in particular, syntax – and aimed to show that some reasonably powerful computational machinery was not up to the task of representing it, before going on to propose a more powerful theory that could. Rosenblatt took some very simple computational machinery – mathematical analogues of neural activation and synaptic connections – and aimed to show that it could represent "low level" cognitive processes involved in object perception and recognition, and that these representations could be learned algorithmically.

Although Chomsky and Rosenblatt never met (Chomsky, e-mail 11/17/17), and neither one's research seems to have had an impact on the other, both of them interacted with members of the intellectual community working on what was dubbed Artificial Intelligence (AI) in 1956 by John McCarthy and Marvin Minsky.[2] Chomsky's arguments about the representational complexity of language were made in the context of the models being explored by that community, in particular Finite State Automata, and related probabilistic Markov chains. George Miller gives the date of a presentation of that work at an AI conference, September 11, 1956, as the birthdate of cognitive science (Miller 2003:142). Rosenblatt presented an early version of his perceptron research at MIT to an AI group in the fall of 1958 (Jack Cowan in Anderson & Rosenfeld 2000:99–100). Amongst the members of that audience was Minsky, Rosenblatt's high school colleague, and the future first author of the critical appraisal of perceptrons mentioned above. Section 2 discusses how Chomsky and Rosenblatt's proposals each diverged from "mainstream AI".

Generative linguistics and neural network modeling developed in apparently complete isolation from one another until they collided thirty years later, when Rumelhart and McClelland (1986) developed a perceptron-based, or connectionist, model of past tense formation in English, which was fiercely criticized from a linguistic perspective by Pinker and Prince (1988) and others. The broader debate between proponents of "algebraic" approaches to cognition, like generative linguistics, and "analogical" models, like connectionism, defined much of the landscape of cognitive science as it developed in the late 1980s and early 1990s. The second section of the paper discusses some of that debate, and argues that it produced important lessons for future research in both traditions, rather than ending in victory for one or the other side. I also survey some characteristics of each of the paradigms, and point out that although the characteristics can usefully differentiate them, their use in contrastive definitions can also lead to false dichotomies.

[2] Jordan (2018) provides a critical discussion of current uses of the term AI, and also makes the interesting historical point that while current AI is dominated by neural and statistical approaches, McCarthy coined it to distinguish his preferred logic-based approach from Wiener's earlier cybernetics, which was more statistical.

The final section of the paper surveys some of the subsequent research over the last thirty years that has integrated aspects of neural network modeling research and generative linguistics. One instance of such integration is Optimality Theory (OT; Prince and Smolensky 1993/2004), whose founders were on opposite sides of the debates of the 1980s. I also discuss the recent resurgence of interest in neural networks in AI, and the emergence of renewed study in cognitive science and AI of their ability to represent linguistic generalizations, with or without explicitly coded linguistic representations.

Based on this survey, I argue that progress on a core goal of generative linguistics, the development of a theory of learning, may well be aided by its integration with neural modeling. The principles and parameters framework (Chomsky 1980; see sec. 3.1 below) is built upon the premise that specifying a universal set of linguistic principles and parameters that delimit the range of possible linguistic variation helps to solve the logical problem of language acquisition. In the absence of a specified theory of learning, however, the argument is not completely solid that a rich Universal Grammar (UG) of this type – or its OT equivalent – is necessary to explain language acquisition. With the development of the rich theories of learning represented by modern neural networks, the learnability argument for a rich UG is particularly threatened. The question of how much and what kind of explicitly pre-specified linguistic structure is needed to explain language acquisition is in fact now receiving renewed attention in light of the learning capabilities of current neural networks. From a review of this work, it is hard to escape the conclusion that a successful theory of learning from realistic data will have a neural component. It is much less clear that a successful theory will need pre-specified UG parameters or constraints, though it seems likely that structured representations like those assumed in generative linguistics will play a role. All of this is discussed in more detail below.

One reason for the general gulf between generative linguistics and neural network modeling may be that research in both traditions can be impenetrable to an outsider. My hope is that this paper will contribute to bridging that gulf not only by giving reasons to build the bridges, but also by providing a relatively accessible introduction to each tradition, inasmuch as this is possible in a single paper, rather than a pair of books.

2. Foundation

2.1 Neural networks

There have been three waves of research on neural networks: in the late 1950s and early 1960s, in the 1980s, and now in the 2010s (see Elman et al. 1996; and Marcus 2001 for introductions to neural nets for cognitive science). Frank Rosenblatt was the most prominent representative of the first wave. His research attracted the attention not only of other scientists, but also of the media,[3] and perhaps partly because of the media attention, it became the focus of considerable controversy, culminating in the publication of Minsky and Papert's (1969/1988) critique (see Olazaran 1993; 1996 for a thorough discussion of the (social) scientific history; and Nagy 1991 for a useful concise summary of research by Rosenblatt and his group).

[3] The New York Times' July 7, 1958 article about a perceptron demonstration starts with "The Navy revealed the embryo of an electronic computer that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

Although neural network research often involves advanced mathematics, the fundamentals only require simple multiplication and addition. Rosenblatt's (1957) perceptron uses analogues of neural structure developed in earlier work, in particular by McCulloch and Pitts (1943) and Hebb (1949). The activity of a neuron – also called a node or a unit – is represented as a numerical value, often as 1 or 0, *on* or *off*. This activity is passed along synaptic connections to other neurons. The connections are weighted: each one has a real valued number that is multiplied by the signal it receives from an input node. A given node becomes active when the sum of incoming weighted signals exceeds a designated threshold (this is a step-activation function, rather than for instance a sigmoidal activation function).

The following diagram represents a small perceptron that performs object classification. There are two features "+Black" and "+Star", each of which defines an Input node. In Rosenblatt 1957 *et seq*. these are the A units (for Associative), which would have themselves been activated through connections to S units (for Sensory). The weights on their connections to a single node (an Output, or Response unit) are shown beside the arrows representing the connections: 0.75 and 0.34 respectively. Given an activation threshold of 0.5, only black objects will activate the Output node.

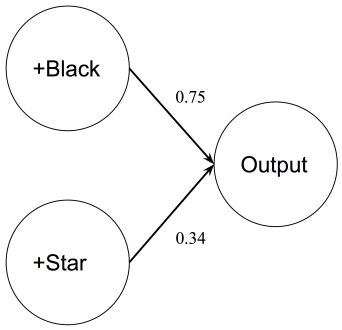

Figure 1: A simple perceptron for object classification

The following table shows this network in action with the four objects instantiating each of the combinations of feature values. Each row shows the object under the Input column, followed by the Input node activations for each of the features. The weighted sum shows the total signal received by the Output node, and its resultant activation is shown in the final column.

(1) *A perceptron classifying the set of black objects*

Input	+Black 0.75	+Star 0.34	Weighted Sum	Activation (> 0.5 input)
★	1	1	1.09	1
☆	0	1	0.34	0
♦	1	0	0.75	1
◇	0	0	0	0

Different positive and negative weights will lead to different sets of objects activating the Output node. Famously, the sets characterized by an "exclusive or" logical relation (XOR) cannot be picked out by this type of network (see Minsky & Papert 1988 for a

proof). For instance, the objects in the middle two rows, the white star and the black diamond, cannot be the only ones to activate the Output node. These two objects form an XOR set: "+Black, or +Star, but not both". One way to get this classification would be to include an Input node that is activated by the conjunction of the features [+Black] and [+Star]. The weight on the connection from that node could then be given a negative value sufficiently high that the black star's activation falls beneath the 0.5 threshold, even while the other black object and the other star's activation are above it. The classificatory power of this simple type of perceptron is dependent on content of the Input nodes: with arbitrarily complex combinations of features, it can perform arbitrarily complex classifications.

In current parlance, this type of perceptron is called a single-layer neural network: there is a single set of adjustable weights connecting the Input layer of nodes to the Output layer (which was in our case just a single node).[4] Besides increasing the size of the Input layer, another way to capture the XOR class is to introduce an additional "hidden" layer of neurons intermediate between the Input and Output, as shown in our next example network, illustrated in figure 2.

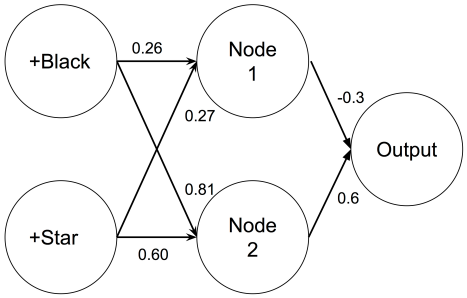

Figure 2. A multi-layer perceptron that performs XOR classification.

The weights on the connections to hidden layer Node 1 from the Input nodes are 0.26 and 0.27. This results in a weighted sum of 0.53 for the black star, and activation only for that object given the 0.5 threshold. This thus implements logical AND. Because the weight from Node 1 to the Output node is negative, the black star receives a penalty, which implements the "but not both" clause of XOR.

The first table below, in (2), shows the mapping from the Input to the hidden layer nodes. The weights to each of them are arrayed vertically, as are the resultant weighted sums and activations. The Node 1 values are bolded. As just discussed, the first node is active only when both features are present, while the second is active for objects that are either +Black or +Star (logical OR).

[4] These are sometimes called two-layer networks (e.g. Pinker and Prince 1988, Marcus 2001), since there are two layers of nodes (Input and Output). This has the downside of making the term "multilayer" opaque.

(2) *Multi-layer perceptron part 1: Input to hidden layer*

Input	+Black **0.26** 0.81	+Star **0.27** 0.60	Weighted Sum	Activation (> 0.5)
★	1	1	**0.53** 1.41	**1** 1
☆	0	1	**0.27** 0.60	**0** 1
♦	1	0	**0.26** 0.81	**0** 1
◇	0	0	**0** 0	**0** 0

The activation of the nodes in the hidden layer, shown in the rightmost column of the table in (2), is the input for the next table, which shows the mapping from the hidden layer to the Output. The activation values (2) are copied into the Node 1 and Node 2 columns of (3). With the assigned weights, the Output node is activated when just Node 2 is active (white star or black triangle), but not when both Node 1 and Node 2 are (the black star).

(3) *Multi-layer perceptron part 2: Hidden layer to Output*

Input	Node 1 –0.3	Node 2 0.6	Weighted Sum	Activation (>.0.5)
★	**1**	1	0.3	0
☆	**0**	1	0.6	1
♦	**0**	1	0.6	1
◇	**0**	0	0	0

This example shows that for a given set of input nodes, a multi-layer perceptron, that is, a network with a hidden layer, can have greater representational power than a single-layer one. Single-layer networks are limited to linearly separable patterns, while as the XOR example shows, multi-layer perceptrons can represent non-linearly separable patterns (separable in the two-dimensional space defined by the two features). For readers familiar with regression models, a useful analogy may be that the crucial node in the hidden layer is acting as an interaction term.

The greater representational capacity of a multi-layer perceptron comes at a price: it is more difficult to train. Rosenblatt (1957; 1958) developed a learning procedure for single-layer perceptrons that is guaranteed to find a set of weights that yields the desired pattern of activation, if such a set of weights exists (Block 1962; Novikoff 1962; Minsky & Papert 1988). As discussed in section 3.1, this procedure does not work with a hidden layer; the development of relatively successful and efficient learning algorithms for multi-layer perceptrons was a primary factor in the sharp increases in research using neural networks in the 1980s and the 2010s.

In their book originally published in 1969, but circulated before that, Minsky and Papert (1988) provide an in-depth study of the representational capacities of single-layer perceptrons, with a primary aim of showing the limits on what this type of perceptron can

represent. The appearance of Minsky and Papert's book coincided with a shift of focus in the mid-to-late sixties from neural networks to other branches of the new discipline of AI, in particular, logic-based ones that manipulate symbolic representations with algebraic operations (see Nilsson 2010 for an excellent history of AI by a participant in this shift). It is hard to know to what degree it was the force of Minsky and Papert's observations and arguments that led to this sea change, or whether it was that neural net research was largely running out of steam given the theoretical and technical barriers that existed at the time (see Olazaran 1993; 1996 for a particularly useful discussion). It is worth emphasizing that Rosenblatt and his group also worked with multi-layer networks, since this is often overlooked (Rosenblatt 1962; Block, Knight & Rosenblatt 1962; Block 1970; Nagy 1991; Pater 2017).

Rosenblatt was trained as psychologist rather than as an engineer, and was committed to using perceptrons to model human psychological properties as realized in the brain, as the following excerpt from his 1962 book emphasizes (p. vi):

> A perceptron is first and foremost a brain model, not an invention for pattern recognition. As a brain model, its utility is in enabling us to determine the physical conditions for the emergence of various psychological properties. It is by no means a "complete" model, and we are fully aware of the simplifications that have been made from biological systems; but it is, at least, an analyzable model.

Despite Rosenblatt's characterization of perceptrons as brain models underlying cognition, research on neural networks did not have much, if any, impact on the emerging field of cognitive science. A 1978 report on cognitive science commissioned by the Sloan foundation to survey the field lists no such research in its bibliography (Keyser, Miller & Walker 1978). Shortly after that report was written, there would be a quite dramatic upsurge in neural network modeling, and at that point, it became a core part of cognitive science. These developments are covered in section 3.

2.2. Foundations: generative linguistics

Chomsky (1957:13) sets up the analysis of a language as a classification problem:

> The fundamental aim in the linguistic analysis of a language *L* is to separate the *grammatical* sequences which are the sentences of *L* from the *ungrammatical* sequences which are not sentences of *L* and to study the structure of the grammatical sequences.

Distinguishing a well-formed sentence from an ill-formed one differs from the object classification examples discussed in the last section in a number of ways. In particular, order in time matters for a sentence, whereas in the examples discussed above, there was no temporal or spatial relationship between the features.[5] A simple example of the

[5] Goldberg (2012) makes use of this property of simple perceptrons to show that baboons' knowledge of legal *vs*. illegal orthographically presented English words can be modeled without reference to word position or linear order, contrary to a claim in the original study (Grainger et al. 2012). Rosenblatt (1967) makes an interesting proposal about how to encode memory through time in a more complex perceptron,

importance of order in syntax is that *The lion sleeps* is a well-formed English sentence, but **Sleeps lion the* is not. There are also dependencies between the form of items that occur at different points in time, as in subject-verb agreement: e.g. well-formed *The lion sleeps* and *The lions sleep* vs. ill-formed **The lion sleep* and **The lions sleeps*.

One way of encoding the difference between these grammatical and ungrammatical sentences is in terms of allowable transitions between words: *the* followed by *lion* is permitted, but not the reverse, and *lion* can be followed by *sleeps* but not *sleep*. Chomsky (1957: 18) shows that "[a] familiar communication theoretic model for language" can encode these sorts of restrictions. This model, a finite state machine, specifies a set of states, along with allowable transitions between them. For example, we can specify that if a machine is in the *lion* state, it may move into the *sleeps* state, but not the *sleep* state.

As the sentence below makes clear – as does the name of the field that he was laying the groundwork for – Chomsky (1957: 13) proposes to classify sentences as grammatical and ungrammatical in terms of whether or not they are generated by a grammar:

> The grammar of *L* will thus be a device that generates all of the grammatical sequences of *L* and none of the ungrammatical ones.

As a generative device, a *finite state grammar* (FSG) can generate an infinite number of sentences using finite resources, a basic criterion for adequacy that Chomsky sets up for a theory of language. It can do this because it allows for loops. Figure 3 provides an FSG for an example from Chomsky 1957, which generates *the man comes*, *the old man comes*, *the old man comes*, *the old, old, man comes* and so on. The loop transition for *old* results in sentences of unbounded length, and thus a set of possible sentences of unbounded size.

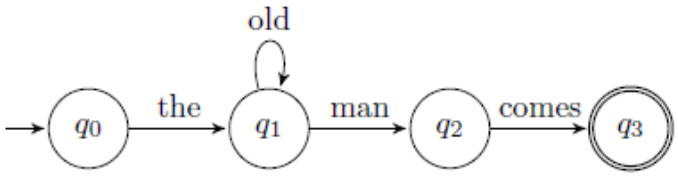

Figure 3. A Finite State Grammar generating an unbounded set of grammatical English sentences

Chomsky (1957) cites Shannon and Weaver's (1949) presentation of information theory as the source of the finite state formalism.[6] While an FSG can represent some

---

one that has some similarities to the recurrent neural networks and convolutional networks used in contemporary deep learning (see section 4.2).

[6] Cherry (1957) provides a useful overview of the state of the art in the modeling of language at the time, with introductions to structural linguistics and information theory; this book was required reading for the first class of MIT PhD students in Linguistics (Barbara Partee, personal communication 2016). In the context of the present joint history of generative grammar and neural networks, it is worth mentioning that Feldman's (1992: 73) encyclopedia entry on finite state machines locates their origin in McCulloch and Pitts' (1943) study of the logical power of mathematical analogues of neurons, and also that Kleene's (1956) introduction of the connection between finite state machines and regular expressions was in the context of "clarifying the results of McCulloch and Pitts" (Block 1970: 513-514).

ordering restrictions and dependencies amongst the words of a grammatical sentence, Chomsky (1957: 21-25) argues that it cannot represent the full complexity of English sentence structure. In particular, Chomsky points to the fact that English allows nested dependencies between particles like *if … then* and *either … or*:

(4) If$_A$ John either$_B$ ate or$_b$ drank then$_a$ he couldn't sleep.
* If$_A$ John either$_B$ ate then$_a$ drank or$_b$ he couldn't sleep.

The grammatical sentence exhibits a mirror structure: if we notate the dependent particle in lower case, the correct structure is *ABba*, and the ungrammatical sentence is instead *ABab*. An FSG cannot generate mirrored structures of unbounded size, just as it cannot represent A^nB^n (see the recent discussion in Jäger & Rogers 2012, as well as the critical commentary on Chomsky 1957 in Pullum 2012). Note that it's not enough for the FSG to encode the fact that an *if* entails a later *then*: it must also keep track of the relative order in which it must encounter the dependent particles. Interestingly, one of the challenges for perceptrons that Minsky and Papert (1988) discuss is the similar recognition of symmetry (*pp.* 117, 252-254). I should note that "mirror structure" is not the conventional term for the linguistic phenomenon, rather, it is usually discussed as "center embedding", and a more typical example would be center embedded relative clauses, such *The girl that hit the balls that flew over the fence is cheering*.

The discussion of the limitations of FSGs in Chomsky (1957: sec. 3) comes after a discussion of the inadequacy of "statistical approximation to English" as a proxy for the grammatical / ungrammatical distinction (sec. 2). This takes as its starting point the famous pair of sentences in (5).

(5) Colorless green ideas sleep furiously.
Furiously ideas sleep green colorless.

Chomsky (1957: 16) claims that since the two sentences, and their subparts, would equally be zero frequency in a corpus, they would be "equally 'remote' from English…in any statistical model of English". Pereira (2000) has shown that this is incorrect: these sentences are in fact distinguished by a bigram model over word categories (note for example that the second sentence starts with a less frequent adverb-noun sequence, compared with the adjective-noun sequence in the first). Nonetheless, Chomsky (1957: 24) may still be right that it would be a "dead end" to try to reduce sentence wellformedness to *n*-gram probabilities over sequences, or to the more complex distributions over sequences that can be represented by a probabilistic Markov chain instantiation of an FSG, regardless of whether those sequences are of words or categories.

So what characteristics must an adequate grammatical model for English have? One obvious characteristic of most grammatical models that goes beyond mere sequential restrictions is the ability to represent the hierarchical phrase structure of a sentence. Chomsky (1957: 30) shows that a phrase structure grammar is able to generate the nested dependencies discussed above, and is thus more powerful than an FSG. Chomsky's proposal goes further than phrase structure in also making use of transformations, which take as input a string with phrase structure and produce as output another string with a new structure (p. 44). Chomsky argues that by deriving passive sentences, negation, and

questions through transformations on a base "kernel" phrase structure, considerable simplifications in the form of the grammar can be obtained.[7]

The relatively deep derivations entailed by a transformational grammar were seen as a liability by some, and alternative frameworks emerged that eschewed them (e.g. GPSG Gazdar et al. 1985; HPSG Pollard & Sag 1994; and LFG Bresnan 2001). Deep derivations have generally been retained in the frameworks that Chomsky himself has subsequently developed, including current minimalism, although the objects being manipulated by the derivations have changed (for a useful recent comparison of minimalism with two non-derivational alternatives, LFG and HPSG, see Bond et al. 2016).

The postulation of these two kinds of abstract structure – hierarchical constituents, and underlying forms that are derivationally transformed into the surface structure – are also characteristic of generative analyses of aspects of language other than syntax. Even though phonological restrictions seem to be representable at the segmental string level by FSGs (Heinz & Idsardi 2011; Heinz & Idsardi 2013), hierarchical representations are standardly viewed as necessary for an adequate characterization of the phonologies of the world's languages (Selkirk 1981; Yu 2017). A well-known example of derivational transformation in phonology is the postulation of an abstract underlying /ai/ in words like *title* and *writer* in Canadian English, in which they are pronounced with a 'raised' diphthong [ʌi] that contrasts with surface [ai] in words like *bridle* and *rider*. The two surface phones can be generated from a single underlying phoneme by assuming that as in the spelling, *title* and *writer* have an underlying /t/, which conditions raising before it becomes the surface flap (Harris 1951; Chomsky 1964). As in syntax, the derivational depth of many analyses in generative phonology has been controversial (see Anderson 1985 for an overview of this controversy).

In phonology at least, one reason that deep derivations are controversial is that they are suspected to pose difficulties for learning (see Dresher 1981 for discussion). Here we can draw a potentially useful connection to the hidden layers of neural nets, discussed in the last section. The perceptron analysis of XOR pattern classification made use of a hidden layer node that abstractly specified the conjunction over features: it was not present in the "surface" Input nodes, which were activated by single features. Like the derivations and hierarchical structure of a generative linguistic analysis, the weights that lead to the conjunctive activation of the hidden layer have to be inferred by a learner. Tesar and Smolensky (2000) call linguistic structure that is not apparent to the learner "hidden", and discuss the learning challenges that it poses (see further 4.1 below). Drawing a connection between hidden linguistic structure and hidden layers of neural nets is potentially useful in two ways, on which I'll expand below. The first is that techniques for learning with hidden layers are potentially useful in learning with explicitly encoded hidden linguistic structure. The second is that hidden layers may be used to learn representations that take the place of explicitly encoded hidden structures. As a historical note, we can also draw a parallel between skepticism about hidden linguistic structure based on learning concerns, and Minsky and Papert's (1988) skepticism about the tractability of learning with hidden layers.

[7] Chomsky's simplicity comparisons are relatively informal. Stabler (2013) presents a related more formal comparison of grammatical succinctness across derivational and non-derivational frameworks. (Thanks to Thomas Graf and Fred Mailhot for discussion).

It is this postulation of abstract structure, specific to language, that differentiates work in generative grammar from what one might call mainstream AI, which tends to postulate relatively concrete representations and domain-general cognitive mechanisms. As we will see in the discussion of connectionism in the sections that follow, generative linguistics is however similar to mainstream AI of the pre-connectionist era in its use of algebraic operations over symbolic representations.

It is perhaps worth reminding ourselves that even the relatively concrete features of a surface linguistic representation, as well as the features used in the neural net discussed in the last section, are themselves abstractions from raw perceptual experience. At a minimum, a learner must acquire a mapping from a more basic level of representation to these features (insofar as it is not prewired), and at least some of these features probably need to be induced. Given the usefulness of neural networks in vision and speech recognition applications in AI, they may well be useful in addressing this part of the learning problem in cognitive science. I will have no more to say about it here though, as my focus is on the "higher level" learning problems that are the focus of generative linguistics.

3. Friction

3.1 Innatism and emergentism

One of the main themes of this section is a caution against the false dichotomies that can be created by contrastively labeling neural network and generative linguistic research, so it is with some trepidation that I start this subsection with a heading that contains two of those very labels. However, to finish setting the stage for the encounter between the generative and neural network traditions that occurred in the 1980s, it is important to emphasize how central learning had become in generative linguistics, and how dramatically the approaches to learning differed in the two traditions.

In Chomsky 1957, there is no real mention of learning, though his future emphasis on learnability considerations is foreshadowed in two ways. The first is in the aforementioned discussion of the inadequacies of (probabilistic) sequential models of language, which can be trained by relatively simple learning algorithms. The second is in a section on the "Goals of Linguistic Theory", in which he rejects the structuralist insistence on discovery procedures – algorithms for proceeding from a corpus to an analysis – in favor of a weaker requirement that there be an evaluation procedure, a method for choosing amongst hypothesized grammars. In Chomsky 1957, the evaluation procedure is for the linguist's task of choosing amongst analyses and theories, but it would later be taken to be part of the acquisition process (Chomsky 1965; Chomsky & Halle 1968).

Shortly after the publication of *Syntactic Structures*, learning did of course become an explicit focus of Chomsky's attention, in his 1959 review of Skinner's (1957) *Verbal Behavior*, which contains the following passage (Chomsky 1959: sec. V):

> It is often argued that experience, rather than innate capacity to handle information in certain specific ways, must be the factor of overwhelming dominance in determining the specific character of language acquisition, since a child speaks the language of the group in which he lives. But this is a superficial argument. As long as we are speculating, we may consider the possibility that the

> brain has evolved to the point where, given an input of observed Chinese sentences, it produces (by an induction of apparently fantastic complexity and suddenness) the rules of Chinese grammar, and given an input of observed English sentences, it produces (by, perhaps, exactly the same process of induction) the rules of English grammar; or that given an observed application of a term to certain instances, it automatically predicts the extension to a class of complexly related instances. If clearly recognized as such, this speculation is neither unreasonable nor fantastic; nor, for that matter, is it beyond the bounds of possible study. There is of course no known neural structure capable of performing this task in the specific ways that observation of the resulting behavior might lead us to postulate; but for that matter, the structures capable of accounting for even the simplest kinds of learning have similarly defied detection.

From this initial speculation about rapid induction, the generative program eventually became one of mapping out the hypothesis space that a learner was claimed to deductively navigate in acquiring a language – of characterizing Universal Grammar (UG).[8] In principles and parameters theory (Chomsky 1980: 3–4), the hypothesis space is characterized by a set of universal principles, alongside language-specific parameters that are "fixed by experience". The postulation of a relatively rich innate endowment is justified in learnability terms: "UG must be sufficiently constrained and restricted in the options it permits so as to account for the fact that each of these grammars develops on the basis of quite limited experience". Chomsky's claims about the "poverty of the stimulus" have engendered considerable discussion (see e.g. Pullum & Scholtz 2002 and the replies in the same volume, and the comments in Pereira 2000: 1243-1245 from an information theoretic perspective, as well as the recent connectionist proposals in Fitz and Chang 2017, discussed further in 4.2 below). One fundamental issue, to which I will return in the conclusion, is that Chomsky's claim about learnability in the just-cited passage was made in the absence of a specified theory of learning.

While Chomsky's argument for a restrictive theory of UG was made on the basis of learning considerations, a restrictive UG-based theory of language typology has been taken as an independent goal in much subsequent generative research in both syntax and phonology. Because of this focus on restrictiveness, generative critiques of connectionist models would thus not only point out their failures to acquire linguistic systems, but also their ability to learn patterns that fall outside those attested in human language (see 3.2 below).

Principles and parameters theory led to a large body of UG-based typological research, as well as considerable research on first and second language acquisition, and to some extent theories of learning addressing the question of how parameters are set (see e.g. papers in Roeper & Williams 1987; Lidz, Snyder & Pater 2016). The parameter-setting problem is non-trivial, and I will return to it in section 4.1 in the context of alternative models of linguistic typology.

At about the same time as principles and parameters theory emerged, neural network research was making large strides in the development of algorithms for the

[8] The extent to which the hypothesis space is specified varies across proposals about UG. Principles and parameters theory, and Optimality Theory after it, have relatively richly specified UGs, even in comparison to many other generative approaches. See sections 4 and 5 below for related discussion.

navigation of the representational space provided by multi-layer perceptrons. Rosenblatt's (1957, 1958) learning algorithm for single-layer perceptrons is an error correction procedure. Given an input, the network is used to predict an output. If the network's output fails to match the correct output, yielding an error, the weights are changed slightly in the direction of generating the correct activation pattern. The activation pattern of a hidden layer is not given as part of the training data, so it is not straightforward to update its input weights. Rosenblatt could see that a solution would be to propagate the error signal back through the hidden layer (and even used the term backpropagation; Rosenblatt 1962: 297-298, cited in Olazaran 1993: 391) but neither he nor any of his contemporaries could find an effective way of doing so. According to Terrance Sejnowski (interview cited in Olazaran 1993: 398), the crucial step in developing the algorithm known as backpropagation (Werbos 1982; Rumelhart, Hinton & Williams 1986; LeCun 1988) was to replace the step activation function, which yields discrete activation levels, with a continuous sigmoidal function. This allowed for the calculation of a gradient, which determines the direction of the weight updates. Although back propagation is not guaranteed to find an optimal set of weights, it is (perhaps surprisingly) effective, and a variety of methods exist for increasing the likelihood that it will find a global, rather than a local, minimum of error.

Because of its focus on learned representations, neural network research is a largely emergentist tradition, and the connectionist linguistic literature often contrasts itself with Chomskyan innatism (Elman et al. 1996; Bates et al. 1998). It is important to emphasize, though, as the just-cited authors do, that a network needs a specified structure for it to represent the effects of learning, just as innate parameters need a specification of how learning works if they are to respond to experience.

3.2 The past tense debate

Even though it was the development of backpropagation for learning with hidden layers that launched the revitalization of neural network research in the 1980s, the model that became the focus of the debate between generativists and connectionists was a single layer network, trained by a version of Rosenblatt's (1957) perceptron update rule. The Input to this network is a phonological representation of the uninflected form of an English verb, and its Output is the predicted phonological form of its past tense. Rumelhart and McClelland (1986:217) present this network as illustrating an alternative to the views that "the rules of language are stored in explicit form as propositions, and are used by language production, comprehension and judgment mechanisms" and that in learning, "[h]ypotheses are rejected and replaced as they prove inadequate for the utterances the learner hears", using a learning mechanism that has "*innate* knowledge of the possible range of human languages". That is, they set up their past tense simulation as illustrating an alternative to the generative approach to language and its learning (for which they cite Pinker 1984 as a state-of-the-art example).

Rumelhart and McClelland's model of past tense formation represents an alternative to the generative view in the sense that there is neither an explicit rule that adds a past tense morpheme to an uninflected stem, nor are there explicit rules for determining the phonological shape of that morpheme. In fact, there is no morphology at all in the model besides the fact that the Input is itself an uninflected stem; there is no morphological decomposition in the Output past tense or in any intermediate form. Both

the addition of the phonologically appropriate form of *-ed* in regular past tense formation, as well as the various forms of irregular past tense such as vowel change (*sing*, *sang*), no change (*hit*, *hit*), and suppletion (*go*, *went*) are all handled by a single network of weighted connections between the atoms of the phonological representation of the uninflected form and those of the inflected one. Learning consists of weight adjustments in response to errors in the prediction of the past tense form. This may be seen as a form of hypothesis testing in the sense that the current values of the weights represent the network's current hypothesis, which is modified by a weight update, but it is different from most generative learning algorithms in that the hypotheses are over a continuous rather than a discrete space, and the changes in output are typically gradual, rather than abrupt (see Elman et al. 1996: ch. 4 on nonlinearities in learning curves). Rumelhart and McClelland (1986) show that these gradual weight adjustments allow them to model the trajectory of the acquisition of the past tense, producing the U-shaped development that results from initial accurate production of irregulars, followed by over-regularization (*goed, hitted*), and then back to the correct target form.

One of the targets of Pinker and Prince's (1988) critique of Rumelhart and McClelland 1986 is the nature of the phonological representations used in the model. Like the Input layer of the object classification networks in section 2 above, there is no temporal order in the Input nodes of the past tense model (see Elman 1990 for a discussion of the difficulties with a temporally ordered Input layer). To represent phonological contexts, a single node encodes features of both the preceding and following phone, as well as the central phone – a triphone representation called a Wickelfeature (after Wickelgren 1969). This is clearly not a general solution to the problem of encoding temporal order, and as Pinker and Prince (1988) discuss at length, it runs into number of problems. Amongst these is the fact that it can represent string reversals (*mad* → *dam*) as easily as it can represent an identity map, yet no language uses string reversals as a phonological process. Subsequent connectionist models of the past tense, as well as other types of morphophonology (see Alderete & Tupper 2018 for an overview), tend to adopt one of two general approaches. One is to notate features for where they appear in a word, as in the templates proposed by Plunkett and Marchman (1993). Another is to make use of a Recurrent Neural Network (Elman 1990), in which the phonological string is processed one segment at a time (see further sec. 4.2). Touretzky and Wheeler (1991) directly address the overgeneration problem noted by Pinker and Prince (1988) by developing a connectionist model that performs mappings from one string to another, and which cannot represent reversals (see also Gasser & Lee 1990 on the difficulty of reversals in Simple Recurrent Nets).

Pinker and Prince (1988) question several other aspects of the representation of past tense formation in the Rumelhart and McClelland model. The critique that most defined the course of future research was the claim that the irregular and regular past are the product of separate systems, rather than being produced by a single cognitive module. In Pinker and Prince's view (p. 122-123), regularities in "irregular" past tense formation, such as the association of particular expressions of the past tense with particular final consonants (e.g. [ei] to [ʊ] with a final [k] as in *take/took, shake/shook*, no change with final [t] or [d] as in *hit, bid*), are "family resemblances" that are the product of the memory system, which may well be formalized in connectionist terms. That is, the irregular past tenses are lexically stored, and anything that looks like a rule-governed

regularity is in fact a product of how the words are stored. The regular pattern, on the other hand, is the product of a morphosyntactic rule, or rules, that add the *-ed* morpheme, and phonological rules of voicing assimilation and vowel epenthesis that yield the contextually appropriate surface forms. For overviews of the subsequent research from what we might call the Pinkerian perspective, see Pinker (1999), Pinker and Ullman (2002) and Marcus (2001: 68 ff.), and from the connectionist perspective, see McClelland and Patterson (2002) and Seidenberg and Plaut (2014). Debates between proponents of one and two systems models are pervasive in cognitive science; for a recent trenchant critique of two systems models of category learning, see Newell et al. (2011).

Since I have been presenting Pinker and Prince's critique as coming from a generative perspective, I should be clear that a two systems approach is not inherent to a generative analysis of the past tense, and is probably not even standard in that tradition. Chomsky (1957: fn. 8) in fact sketches a rule-based analysis of cases of vowel change like *take/took*, which is developed in Chomsky and Halle (1968: p. 11) and Halle and Mohanan (1985: 104ff.), and Albright and Hayes (2003) present a quite different single system rule-based analysis. Although many phonologists might in fact believe that lexically irregular morpho-phonology is better handled by a system of lexical analogy than by a rule-based system (see Albright and Hayes for a critique of that view), it is rare to see that view developed into analyses, presumably because most phonologists do not have a formal system at hand for constructing the lexical part of the analysis. A notable exception is Bybee's model of Natural Generative Phonology, which draws a strict distinction between productive phonological rules and semi-productive morpho-phonology; Bybee (1988) embraces connectionism as a means of formalizing her lexical networks (see further Bybee & McClelland 2005).[9]

The debate between connectionists and proponents of rule-based models of cognition often turns on a definition of what it means for a model to be rule-based, or connectionist. Lachter and Bever's (1988) critique of Rumelhart and McClelland (1986), and Marcus' (2001) critique of its connectionist successors, challenge the extent to which the models can truly be claimed to be alternatives to rule-based models. Although Rumelhart and McClelland's model of the past tense clearly lacks the rules of a standard analysis of the regular past tense, it adopts relatively standard phonological features, and Lachter and Bever (p. 211) argue that choice, as well as the particular configurations of features for the nodes, essentially engineers a rule-based solution into the model. Marcus (p.83) argues that:

> the closer the past tense models come to recapitulating the architecture of the symbolic models–by incorporating the capacity to instantiate variables with instances and to manipulate…the instances of those variables–the better they perform.

[9] As Bermúdez-Otero (2016) points out, another generative approach that seems compatible with the two-systems view is Jackendoff's (1975) work on Lexical Redundancy Rules, and there are also relations to ideas in Lexical Phonology (Kiparsky 1982; see Kaisse & Shaw 1985; Kenstowicz 1994 for overviews). See also Liberman (2004) on the divergence between generative practice and Pinker's two systems approach, and Embick and Marantz (2005) on some of the issues in relating generative theories to experimental data on the past tense.

The absence of variables in the Rumelhart and McClelland (1986) model is another of the primary targets of Pinker and Prince's (1988) critique (see p. 176), and it is the presence of variables that Marcus (2001) takes as part of the definition of a "symbolic" model, and key to its success – the other part is the ability to manipulate those symbols with algebraic rules (see Smolensky 1988 on the subsymbolic nature of connectionism; and see Hummel 2010 for a recent discussion of the properties of symbolic systems).

Some of the back-and-forth in the past tense debate can be tiring precisely because it consists of one side accusing the other of not being true to its principles in incorporating aspects of the first sides' theory. But this aspect of the debate is ultimately instructive in that it shows that the space between connectionist and generative models of language is more fluid than the rhetoric might sometimes suggest. Some points related to this fluidity have already been made above: (1) there is nothing about a connectionist model that prohibits the use of symbols, including variables, and other representations developed in linguistic traditions (see e.g. Doumas, Puebla & Martin 2017; Palangi et al. 2017 for recent work on learning symbolic representations in neural nets); (2) a generative rule-based model can, and often does, have the very specific rules needed to model irregular morpho-phonology; (3) a generative model is not fully innatist in that parameters need to be set by experience; (4) a connectionist model is not fully emergentist in that much of its structure must be specified. None of this is controversial, but it might not be apparent when each of the traditions is contrastively labeled with these characteristics.

Before moving on to discussion of contemporary models that illustrate the fruitfulness of integration across the connectionist-generativist divide, it is worth saying a few words about another model characteristic that might be taken as definitional of each of generativism and connectionism. Rumelhart and McClelland's (1986) past tense model uses a probabilistic interpretation of a sigmoid activation function, and thus produces probabilities over different outputs for a given input. Models of generative grammar, from Chomsky (1957) onwards, typically use deterministic rules that produce a single output for a given input.[10] Neither of these choices are fixed however: the examples of perceptron networks in section 2 are deterministic, and rules can be given a probability of application (see e.g. the probabilistic formulation of Labov's variable rules in Cedergren & Sankoff 1974; and the probabilistic formulation of minimalist syntax in Hunter & Dyer 2013; see Lau, Clark & Lappin 2017 for recent discussion of probabilistic grammars and sentence acceptability judgments). The past tense debate also provides a reason to formalize a rule-based model probabilistically: as children acquire the regular *-ed* past tense, its probability of use seems to increase gradually (McClelland & Patterson 2002:467–468).[11] Pinker and Prince (1988:164) in fact sketch a rule-based approach to acquisition that incorporates "competition among multiple regularities of graded strength", which is elaborated on with explicitly stochastic rules in Albright and Hayes (2003).

[10] See the interview with Chomsky in Katz (2012) for an expression of continued skepticism about probabilistic approaches (to AI), and Norvig (n.d.) for an extended reply to earlier related comments.

[11] A two systems theory might not in fact need probabilistic rules to capture these data, provided it had probabilistic lexical access that competed with rule application. As far as I know, there are no implemented versions of such an approach that compete with the connectionist accounts (see Zuraw 2010 for a related proposal).

In my view, rather than yielding a single victor, the past tense debate provided important lessons for the further development of both the generativist and connectionist traditions, as well as their hybrid offspring. The connectionist models of the past tense, including that of Rumelhart and McClelland, show that simple and explicit models of learning can be combined to good effect with the representational structures developed in linguistics: learning need not be over unstructured strings of words, as in the models criticized by Chomsky (1957). The generativist critiques bring to the fore the structural complexity of language, which is not well captured by Rumelhart and McClelland's model or its immediate successors. Perhaps more controversially, they also seem to indicate the fruitfulness of incorporating richer combinatorial structure in the representations manipulated by connectionists' models.

4. Fusion

Despite – or perhaps because of – the debates between proponents of connectionist and generative approaches to the study of language, at the end of the 1980s and beginning of the 1990s some linguists began developing connectionist models of language, which often incorporated representational assumptions from generative linguistics (e.g. Lakoff 1988; Hare, Corina & Cottrell 1989; Legendre, Miyata & Smolensky 1990; Goldsmith 1993; Lakoff 1993; Wheeler & Touretzky 1993; Gupta & Touretzky 1994). Interest developed from the other "side" as well: some of the just cited papers were collaborations with connectionists (Smolensky, Touretzky), and connectionist research on language often began to make use of explicitly hybrid models (see e.g. many of the contributions to Sharkey 1992).[12] In the following sections, I cover a small part of the work[13] over the last 30 years that synthesizes aspects of connectionism and generative linguistics: section 4.1 discusses an approach to generative grammar resulting from the importation of constraint interaction, a relatively high-level abstraction from connectionism, while section 4.2 goes on to discuss the ongoing controversy over the representation of syntactic structure in connectionist networks.

4.1 Constraint interaction in generative grammar

The title of this subsection is the subtitle of Prince and Smolensky's 2004 book originally circulated in 1993 that introduced Optimality Theory (OT), a particularly fruitful connectionist-generativist fusion (see the now more than 1300 contributions to the Rutgers Optimality Archive – http://roa.rutgers.edu).

OT is a descendent of Chomsky's (1980) principles and parameters framework in that it similarly posits a rich UG, with the goal of delimiting the space of possible languages. Instead of parameters whose values are fixed by a learner, OT has constraints whose ranking must be determined. A close relative of OT is Harmonic Grammar (HG; Legendre, Miyata & Smolensky 1990; Smolensky & Legendre 2006), which numerically

[12] Rosenblatt was himself an advocate of a hybrid approach to cognition, in the sense that he saw some cognitive processes as outside of the domain of perceptron theory (Rosenblatt 1964).

[13] One particularly notable omission is research in the Gradient Symbol Systems framework (Smolensky, Goldrick & Mathis 2014), which would require another paper to cover. I have also in general abstracted from the question of whether gradient well-formedness provides arguments for connectionist approaches (see Legendre, Miyata & Smolensky 1990 on the modeling of gradient syntactic judgments with weighted constraints).

weights, rather than ranks, its constraints. I will illustrate the OT/HG formalization of generative grammar using HG's weighted constraints, since this represents a more direct fusion with connectionism (see Pater 2016a for discussion of HG/OT similarities and differences).

Prince and Smolensky (2004; ch. 4) provide an extended argument for constraint interaction in the domain of word stress, comparing the Extrametricality parameter (Hayes 1980) to a violable Nonfinality constraint. An Extrametricality parameter places word-final syllables outside of the domain of word stress, while a Nonfinality constraint penalizes stressed word-final syllables, but can be violated under the compulsion of other constraints. To see how violable constraints can produce different results from inviolable rules or constraints, consider the interaction of Nonfinality with a Weight-to-Stress constraint that requires heavy syllables to be stressed. These constraints come into conflict when a heavy syllable is in final position. The tableau in (6) shows a bisyllable with a heavy final syllable (one with a coda nasal) and a light initial one (a CV syllable), and two candidate stress placements. The first candidate, *batán*, has stress on the final syllable, and thus violates Nonfinality, indicated with a negative integer in its column. The second candidate has stress on the light syllable, so the unstressed heavy violates Weight-to-Stress. The constraints' weights are provided underneath their names: Weight-to-Stress has a higher weight than Nonfinality (5 *vs*. 2). The column labeled Harmony gives the weighted sum of constraint violations. In a deterministic version of HG, the candidate with highest Harmony, in this case *batán*, is picked as the Output – the optimal candidate.

(6) *A Harmonic Grammar tableau*

Input: batan	Weight-to-Stress 5	Nonfinality 2	Harmony
Output: batán		–1	–2
bátan	–1		–5

The notion of constraint interaction – of one constraint overriding another – illustrated in this tableau can be abstractly seen in the connectionist network for XOR in section 2.1, in which the constraint against black stars activating the output neuron (negative weight on the relevant connection) outweighs the constraint that black objects and stars should activate it (positive weights on the relevant connection).

So far, we could equally analyze this situation with an Extrametricality parameter that is turned off, so that the final syllable is eligible to be stressed, and a Weight-to-Stress parameter that is turned on, so that the heavy syllable is picked over the light one. Consider, however, the following tableau, which shows a form in which there is no heavy syllable, and in which Nonfinality can be satisfied without violating Weight-to-Stress. Nonfinality prefers the candidate with stress on the initial syllable over stress on the final syllable. If Nonfinality were replaced by an inactive Extrametricality parameter, it could not be used to account for the lack of final stress in this instance. Cases like this are referred to in the OT literature as the emergence of the unmarked, or more broadly, as non-uniform constraint application.

(7) *Illustration of "emergence of the unmarked"*

Input: bata	Weight-to-Stress 5	Nonfinality 2	Harmony
batá		–1	–2
Output: báta			0

The switch from parameters to violable constraints has consequences for both the study of language typology, and of learning.[14] For language typology, it becomes possible to maintain relatively general formulations of constraints while still accounting for details of individual languages. In parametric theory, an analyst's observation of surface violations of constraints in a language in which they are generally active leads to a range of responses, including simply changing a general constraint into a set of more specific ones that can remain inviolable – this tactic is the main target of Prince and Smolensky's (2004) attack on Extrametricality (see also Coetzee 2008: sec. 2.1 on the violability of the OCP). In a review of Halle and Vergnaud's (1987) parametric metrical theory of word stress, Dresher (1990) discusses some of its additional "extraparametric devices", and concludes that their use draws into question Halle and Vergnaud's claims that their theory is more typologically restrictive than its competitors. Dresher (p. 184) is in fact pessimistic that the relative restrictiveness of the theories can be determined: "At a time when all versions of metrical theory command such arsenals, comparisons of expressive power are likely to remain inconclusive."

The situation changed dramatically with the introduction of violable constraints (see e.g. Kager 2005 for an extended comparison of two OT theories of word stress). The greater success in constructing and comparing theories of typology was partly due to the success of the associated learning algorithms. Both OT and HG have provably convergent learning algorithms that will find a constraint ranking, or weighting, for any set of candidate outputs that can be made jointly optimal (see Tesar & Smolensky 2000 on OT; Potts et al. 2010; Boersma & Pater 2016 on HG). Given the candidate sets and their violation profiles, we can thus determine which potential languages (sets of optima) are in fact generated by the constraint set. This learnability approach to typology calculation was pioneered by Hayes et al. (2013) for OT (see also Prince, Tesar & Merchant 2015), and has been extended to HG and serial variants of OT and HG by Staubs et al. (2010).

Studies of typology in OT and HG typically make use of deterministic variants of the theories, in which within-language variation is abstracted away from. Probabilistic variants of OT and HG have also been developed (see the survey in Coetzee & Pater 2011), and alongside them learning algorithms. Maximum Entropy Grammar (MaxEnt; Goldwater & Johnson 2003), a probabilistic variant of HG, has the distinction of having associated provably convergent learning algorithms, due to its grounding in

[14] I am abstracting from another dimension on which OT and parameter-based theories differ. Prior to OT, constraints in generative grammar interacted with rules, rather than with one another (see McCarthy 2002; Prince & Smolensky 2004; McCarthy, Pater & Pruitt 2016 for comparison). The story of this particular marriage between connectionism and generative linguistics is largely about the quest for improved constraint-based theories of linguistics – see e.g. the comments by Alan Prince in Pater (2016b) – though it is also about finding a neural grounding for symbolic systems (Smolensky & Legendre 2006). Much of the controversy around OT concerns its abandonment of serial rule ordering (Vaux & Nevins 2008).

mathematically well-understood models from other domains, including neural networks (Smolensky 1986; Johnson 2013a). In MaxEnt, the probability of an output is proportional to the exponential of its Harmony. For example, with the weights in Tableau (6), *batán* would have probability 0.95, rather than being deterministically picked as optimal, since $exp(-2) / (exp(-2) + exp(-5)) = 0.95$.

As well as modeling variation in final state grammars, probabilistic OT and HG grammars have also been used to model the variation in grammars in the course of acquisition (see the overview in Jarosz 2010; see also Moreton, Pater & Pertsova 2015). One class of gradual learning algorithm used in this work includes an application of Rosenblatt's Perceptron convergence procedure (Pater 2008; Boersma & Pater 2016); this class of inter-related algorithms also includes Boersma's (1997) widely used learner for Stochastic OT (see also Boersma & Hayes 2001) and (Stochastic) Gradient Descent for MaxEnt (Jäger 2007; Moreton, Pater & Pertsova 2015).

The just-mentioned convergence guarantees of OT and HG learning algorithms come with an important caveat: they apply only when the structure of the learning data is supplied in whole – when all the constraint violations of each learning datum are known. To continue with the stress example, finally stressed *batán* might be analyzed as having a trochaic (left-headed) foot on the final syllable, or an iambic (right-headed) foot that parses both syllables. As shown in (8), each representation incurs distinct constraint violations, since in one case the initial syllable is unparsed, violating Parse-Syllable, while the other falls afoul of Trochee, a constraint demanding left-headed feet. If we supply the learner with only the overt form, it must choose between the two full parses (or assign probability to them).

(8) *An example of a hidden structure problem*

Overt form	Full structure	Parse-Syllable	Trochee
batán	ba(tán)	–1	
	(batán)		–1

While it might not be a problem for an analyst to supply full structures when studying typology or in modeling some cases of variation, there are many cases of linguistic analysis in which one might not be committed to a particular full structure for each piece of data, and would like a learner to find an appropriate grammar. In addition, dropping the idealization of full access to structure is part of moving to a more realistic model of human language acquisition.

Learning with hidden structure (≈ learning with structural ambiguity) is generally accomplished in OT and HG by using the current state of the grammar to pick amongst the full structures for a piece of learning data (Tesar & Smolensky 2000), or to assign a probability distribution to them (Jarosz 2013; Jarosz 2015; Boersma & Pater 2016; Boersma & van Leussen 2017). A MaxEnt version of this general approach (e.g. Pater et al. 2012; Johnson et al. 2015; Pater & Staubs 2013) creates a single vector (row) of constraint scores for a partially structured learning datum by summing over the probability weighted vectors of all of the corresponding full structures (see Staubs & Pater 2016; Nazarov & Pater 2017 for extensions to serial variants of MaxEnt). None of these methods is guaranteed to converge on a global optimum, and their development and comparison is an area of ongoing research (as Boersma 2003 points out, non-convergence

can be an advantage, insofar as it corresponds to human learning difficulties or typological gaps).

The learning methods for MaxEnt models and neural networks are highly overlapping, as was already noted in the discussion of gradual learning algorithms. Both can be learned with gradient-based optimization methods, including Gradient Descent (recall that the gradient indicates the direction of weight change). When the gradient for a neural net with one or more hidden layers is constructed using backpropagation, or when a gradient for a MaxEnt model is constructed with hidden structure (see Staubs & Pater 2013 for a derivation of the gradient), there is no guarantee that these methods will find the best set of weights for the model, in terms of optimizing the fit of the model's predictions to the data. That is, the learner may not find the global optimum, and may instead be trapped in a local minimum of error. Because of the current prominence of deep learning in AI, there is substantial research effort being expended to improve the performance of learners for multi-layer perceptrons, and many of the proposals can be adapted directly to MaxEnt (for example, the one in Neelakantan et al. 2015).

Learners for parametric models are also subject to local optima because of structural ambiguity (Gibson & Wexler 1994). A typical response to this problem, in Gibson and Wexler and elsewhere, is to search for triggers (or cues) in the learning data that unambiguously correspond to the non-default value of each parameter, and then build those into the learner; one might also need to stipulate an ordering on the setting of the parameters (e.g. Dresher 1999). It is generally an advantage of violable constraints that such additional machinery is unnecessary (Tesar & Smolensky 2000), but recent work by Gould (2015) and Nazarov and Jarosz (2017) indicates that probabilistic parametric learners may also succeed without triggers.[15]

One particularly interesting application of MaxEnt to the learning of phonology falls somewhat outside of standard OT/HG frameworks, in two respects. Hayes and Wilson (2008) develop a model of phonotactics that defines a probability distribution over the space of possible words. Unlike standard OT/HG as presented above, there is no mapping from an Input to an Output (or to put it differently, the Input is an undifferentiated "word", rather than a word of a particular phonological shape). Also unlike standard OT/HG, the constraint set is not taken as a given (see Hayes & Wilson 2008:425; Moreton & Pater 2011 for discussion of the consequences for study of typology). Hayes and Wilson propose a method for constraint induction that chooses amongst candidate constraints according to a set of heuristics (see Wilson & Gallagher 2016 for an alternative gain-based approach and comparison with other theories; see Berent et al. 2012 on evidence for the incorporation of variables; and further Berent 2013). The hidden layer of a neural net allows for an alternative to constraint induction – which was hinted at in section 2.1, where we saw a hidden layer node sensitive to feature conjunction – and current research is exploring this alternative (Alderete, Tupper & Frisch 2013; Doucette 2017).

[15] Nazarov and Jarosz (2017) find that Yang's (2002) earlier trigger-free probabilistic learner succeeds at learning only 1 of 23 languages generated from a set of metrical parameters; their own learner succeeds on 22 of them. It is also worth noting that Tesar (2004) presents a learner that can provably cope with structural ambiguity; it works only with deterministic OT (or HG), and cannot be used to model human learning paths.

4.2 Can recurrent neural networks learn syntax?

Recurrent neural networks (RNNs; Elman 1990; Elman 1991) do perhaps surprisingly well in capturing some aspects, including long-distance dependencies, of natural language syntax, and have recently undergone a resurgence of popularity in AI applications of neural networks to language (see Goldberg 2016; Goldberg 2017 for tutorial overviews). It is unclear, however, whether they can fully learn syntactic regularities without the incorporation of hierarchical representations of the type used in generative linguistics, or alternative linguistic structure like dependency marking. In this section, I briefly introduce RNNs as used in Elman's studies (often called Simple Recurrent Nets), provide an overview of some of the research on the learning of syntax using them and other types of RNN. This leads into a quick discussion of current neural network models, and their application to other linguistic domains, including the English past tense.

When applied to a sequence of elements, such as letters, phones, or words, RNNs process one at a time, starting at one edge of the string – typically the left (some modern applications scan in both directions). The network is used to predict the next element in the sequence, and the weights are updated based on that prediction (updates can also be made on more global predictions). As illustrated in Figure 4, when moving onto to the next element in a sequence, the current hidden layer is copied as a context layer to provide an extra set of inputs to the next computation of the hidden layer activations, and the Output. The representation encoded in this copied hidden layer provides a basis for the prediction of upcoming elements based on those encountered earlier – i.e. it is a type of sequential memory.

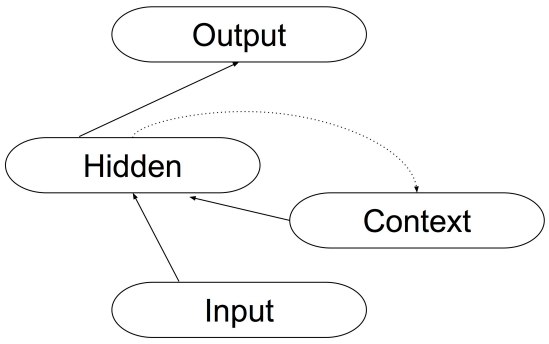

Figure 4. The structure of a Simple Recurrent Neural Network (adapted from Lewis & Elman 2001)

Elman (1990; 1991) uses toy language examples to show that these networks can capture some basic aspects of syntax. Elman (1990) focuses on the formation of categories (Noun and Verb), and on word order restrictions in single clauses. Elman (1991) goes on to examine RNNs' abilities to capture subject-verb agreement dependencies across embedded clauses, finding that their predictions are in fact influenced by information retained from across the embedded clause.

Elman (1991:221) recognizes the preliminary nature of his results, and given the limitations of the datasets he was working with, one could reasonably either be enthusiastic about the potential for further development of this approach, or skeptical. And presumably, that degree of skepticism could well be influenced by the strength of

ones' prior belief in the necessity of a rich UG. From a generative perspective, an RNN is a very impoverished theory of UG. Not only does it lack parameters or constraints encoding the substance of possible grammatical systems, the basic structures, such as syntactic categories and trees, are themselves absent, at least prior to learning. A hidden layer can form abstract representations of the data, and there are some hints in Elman's results that those representations may do the work of explicit categories and constituent structure, but much research remains to be done, even today, to determine the extent to which they can.

Fitz and Chang (2017) provide a particularly useful overview of one line of research in this domain, and some intriguing new results. A much-discussed case of the poverty of the stimulus is auxiliary inversion in English *yes-no* question formation (Chomsky 1975; Chomsky 1980). The fronted auxiliary in *Is the bottle that the woman is holding broken?* is displaced from the underlined verb phrase of the main clause in *The bottle that the woman is holding is broken*, rather than from the embedded clause. This is a structure dependent rule, as opposed to a putative rule of "front the first auxiliary", which would work for mono-clausal cases, but would yield the ungrammatical **Is the bottle that the woman holding is broken?* for this and other similar cases of embedding. Fitz and Chang show that an RNN that is simply trained to predict upcoming words generalizes the incorrect rule (contra Lewis and Elman 2001). They further show that if the learner is given the task of generating sentences given a meaning representation, it does yield the correct generalization from monoclausal training instances of auxiliary inversion to the structure dependent rule. Since they incorporate explicit propositional structure into their meaning representations, Fitz and Chang's work can be seen as an instance of the "fusion" that is the focus of this section – they in fact see their proposal as bridging the usually antagonistic emergentist/constructivist and innatist approaches to aux-inversion (p. 243).

Whether or not Fitz and Chang's approach will generalize to other phenomena is very much an open question. One set of challenges for RNN approaches to syntax appears in Frank, Mathis and Badecker's (2013) study of anaphora resolution. The pair of sentences from their study (p. 197) in (9) illustrates part of the phenomenon.

(9) a. Alice who Mary loves admires herself
b. Alice who loves Mary admires herself

The question is who *herself* refers to. As Frank et al. point out, native speaker judgments, as well as online processing measures (Xiang, Dillon & Phillips 2009), show that it can only be Alice – not Mary, or anyone else. This is captured in grammatical accounts in terms of the positions of Alice and Mary relative to *herself* in a hierarchical structure. Frank et al. explore an RNN similar to that illustrated in Figure 4, except that the context layer is sandwiched between two hidden layers. They train it to predict upcoming words, and then use the trained network to assign interpretations by examining which nodes *herself* activates, where Alice, Mary and Sue each have a node. They find that the network correctly assigns almost all of the probability to Alice in (9a). In (9b.), however, it assigns probability of 0.16 to Mary. They diagnose the success on (9a.) as indicating the ability of RNNs to capture structural relationships (though see further below), and the partial failure on (9b.) as indicating their ability to capture incorrect linear relationships.

The substring *Mary admires herself* can be interpreted in other contexts with herself referring to Mary, but not in (9b.), because of its clausal structure. Importantly, they argue that the sequence-based errors of networks in this task are different from those observed in human sentence processing (p. 200-201), which show sensitivity to structure (though see relatedly Willer Gold et al. 2017).

Neural networks – especially RNNs – without prespecified linguistic structure have recently been broadly applied in AI language tasks with considerable success. In machine translation, RNNs are used to map from a sequence of words in one language to a sequence in another, without any intermediate explicitly encoded linguistic structure (Sutskever, Vinyals & Le 2014; Bahdanau, Cho & Bengio 2016; Wu et al. 2016).[16] These models achieve state of the art performance, doing as well – and usually better – than earlier models that map to intermediate levels with linguistic structure like phrases, and are the basis of useful applications like Google Translate. This "end-to-end" approach has also been applied in speech recognition, mapping from acoustic signal to text without an explicit intermediate phone layer (e.g. Amodei et al. 2016).

The success of modern RNNs in applied language tasks is due to advances in their architecture, as well as in training methods and computational hardware which jointly allow for training of large networks – e.g. very "deep" ones with many layers – on large datasets. Do these advances lead to models that can successfully learn to represent natural language syntax without explicitly specified linguistic structure? Linzen et al. (2016) explore the ability of modern RNNs to learn subject-verb agreement, and find results that are somewhat parallel to those of Frank et al. (2013): the models do have a degree of success on even long-distance dependencies, indicating that they have learned something akin to a structural analysis, but they also have a tendency to extract incorrect linear regularities (see also the follow-up in Bernardy & Lappin 2017). Adger (2017) takes the failures of this model as vindication of Chomsky's (1957) arguments against statistical models of language, but one could equally take its successes as vindication of a statistical approach. Linzen et al.'s own interpretation of their full set of results is that they are encouraging, but that some sort of explicit linguistic structure may need to be added to allow appropriate generalization based on learning data closer to those experienced by humans (that is, under unsupervised learning).

The source of the inappropriate linear generalizations in cases like those discussed in Frank et al. (2013) and Linzen et al. (2016) is not entirely clear. In the limit, RNNs are universal function approximators (see Elman 1991:219 for discussion), so it would be incorrect to say that they cannot represent the correct mappings. Whether a given network can represent a particular set of mappings is another question, as is whether it will wind up in that state with a given learning algorithm operating with a given set of training data. Therefore, further research on the learning of syntax by RNNs – and other types of neural net – will involve exploration of (1) the representational capacities of particular networks,

[16] The RNNs in these sequence-to-sequence models are typically embedded in a Long Short Term Memory architecture (LSTM; Hochreiter & Schmidhuber 1997). Recent work suggests that the LSTM architecture itself, independent of the RNN, may provide much of its representational power (Levy et al. 2018). Other work has suggested convolutional networks as alternatives to RNN/LSTMs (e.g. Bai, Zico Kolter & Koltun 2018). Edelman (2017:105) offers a skeptical assessment of the applicability of Deep Learning (DL) to the modeling of linguistic cognition – "While the practical appeal of DL is clear, its contribution to the understanding of the human language faculty seems limited." – and a survey of alternative neural models that he deems more promising.

and (2) the performance of particular algorithms, (3) and the relationship between properties of the training data and the resulting generalization. In proposing Recurrent Neural Network Grammars, Dyer et al. (2016) hypothesize that RNNs do in fact need to be supplemented with hierarchical structure.

The extent to which explicit linguistic structure is needed in AI models of language is very much a subject of general current debate (see e.g. the presentations and discussions collected in Pater 2018). Alongside the just-mentioned proposal by Dyer et al. (2016), there are a number of recent proposals that incorporate compositional and hierarchical linguistic structures into neural networks with the aim of improving their performance on AI tasks (e.g. Andreas & Ghahramani 2013; Socher et al. 2013; Bowman et al. 2015; Yogatama et al. 2016; though see also Bowman, Manning & Potts 2015). Alongside these proposals are observations about the fragility of models that eschew linguistic structure when they are tested on linguistically challenging data (e.g. Ettinger et al. 2017; Jia & Liang 2017). In her overview of the presentations at the most recent meeting of the Association for Computational Linguistics, See (2017) in fact dubbed one of the main trends as "Linguistic Structure is Back".

The sequence-to-sequence models used in machine translation allow for straightforward applications to the kinds of within-language mappings between representations studied by linguists. These applications have yielded some encouraging results, even for models that do not incorporate explicit linguistic representations. For example, Kirov (2017) and Kirov and Cotterell (2018) train a standard RNN-based sequence-to-sequence model on the same English present-to-past tense mappings as Albright and Hayes' (2003) stochastic rule-based model. They find that its predictions provide a far better match to native speaker formations of novel past tense forms ("wug" test data), especially amongst irregulars, than Albright and Hayes' own model. This provides some reason to be skeptical of Marcus' claim, presented in sec. 3.1 above, that a successful model of the past tense needs symbolic representations.[17] Sequence-to-sequence models can also be used for syntactic transformations, as shown by Frank and Mathis (2007) using an earlier generation of RNN (see also relatedly Chalmers 1990), and ongoing research is addressing the extent to which these models cope with the auxiliary inversion poverty of the stimulus problem discussed above (McCoy, Frank & Linzen 2018).

Besides testing the performance of trained neural networks, one can inspect the values of connection weights and the activation patterns produced for particular inputs to gain insight into the representations they have constructed. Elman (1990; 1991) pioneered this approach with his early RNN studies, finding evidence for the representation of syntactic categories in the hidden layer. In their study of anaphora resolution, Frank et al. (2013) conclude that the representations are not sufficiently abstract, being too tied to particular words rather than to categories. This approach has also recently been applied in

[17] Marcus' (2001:chap. 3) main argument for symbolic models comes from their ability to represent generalized identity functions, and the inability of RNNs to learn these functions (see also Tupper & Shahriari 2016). These identity functions can play a role in representing natural language reduplication. Although initial work found that standard sequence-to-sequence models did not seem to learn general reduplicative mappings (Prickett 2017), further exploration of these models, in particular using the Dropout technique to aid generalization (Srivastava et al. 2014) has yielded positive results (Brandon Prickett p.c.; see also Alhama & Zuidema 2018 for related work).

speech, finding evidence for phone categories being represented in a hidden layer of a network trained to map from acoustic signals to images (Alishahi, Barking & Chrupała 2017). A particularly relevant recent result in this vein with respect to the question of how syntax may be represented in RNNs is presented in Palangi et al. (2017), who introduce Tensor Product Recurrent Networks, and show that their internal representations can be interpreted syntactically. Like the Hayes and Wilson model discussed at the end of the last section, this model occupies an interesting middle ground between the poles of innatism and emergentism, since it is given the structural building blocks of symbols and their roles, but must learn their configurations.

5. Conclusions

When viewed from a sufficient distance, neural network and generative linguistic approaches to cognition overlap considerably: they both aim to provide formally explicit accounts of the mental structures underlying cognitive processes, and they both aim to explain how those structures are learned. When viewed more closely, especially with respect to the research practices within each tradition, they may seem to diverge sharply, with the bulk of connectionist practice involving computational learning simulation allied with AI tasks (see e.g. 4.2 above), or with psychological experimentation (see e.g. papers in Christiansen & Chater 2001), and with the bulk of generative practice involving grammatical analysis of linguistic systems. At a middle depth of field, one can find a growing research territory in which the bodies of knowledge and the models developed in each of these traditions are jointly applicable. In this paper, I have focused on the question of how systems that adequately represent linguistic knowledge can be learned. Some of the most promising avenues for answering that question build both on generative insights into the nature of linguistic knowledge, and on connectionist insights into the nature of learning.

The utility of representations like those assumed in generative linguistics for neural network modeling of language was discussed in sections 3.2 and 4.2. Why might generative linguistics need connectionism for the formalization of learning? At first sight, it is not obvious, insofar as learning in generative linguistics is reduced to the setting of parameters, or to the ranking or weighting of constraints. In section 4.1, I surveyed some ways in which the importation of learning theory from connectionism and related areas of statistical learning has been profitable, and promises to yield further dividends, especially in the context of weighted constraint theories of grammar. The biggest payoff, however, is almost certainly to come in confronting the problem of learning the representations, constraints or rules themselves, as in the work mentioned at the end of both sections 4.1 and 4.2. “Feature induction” is a very difficult problem in AI, and a lot of the success of neural approaches comes from their ability to learn representations of the data in hidden layers (a leading conference for neural net research is the International Conference on Learning Representations).

Learning considerations continue to play a role in current theoretical discussions in the Minimalist framework, just as they did in principles and parameters theory (sec. 3.1 above). For example, in arguing for a parameter-free version of Minimalism, Boeckx (2014) challenges claims that having parameters in UG aids learning, while Chomsky et al. (2017:19) criticize the “Cartographic Program pursued by Cinque, Rizzi and many others” by saying “there is no conceivable evidence that a child could rely on to learn

these templates from experience." As with Chomsky's programmatic statements about principles and parameters theory discussed in section 3.1, it is hard to know how to assess these sorts of claims in the absence of a learning theory. Hunter and Dyer (2013) show that Minimalist grammars can be formalized as log-linear models (i.e. MaxEnt grammars), thus opening the door to the importation of the learning theories discussed in 4.1 (see Johnson 2013b for a more general introduction to statistical learning of grammars). It has been pointed out that statistical learning theory, especially Bayesian modeling, can permit a more rigorous assessment of claims about UG (see the overview in Pearl & Goldwater 2016).[18] When neural network modeling is integrated with grammatical formalisms in the ways discussed in section 4, we may be able to go further in assessing the extent to which grammatical representations can be learned from experience, and what aspects of the grammar must be hard-wired. In developing grammatical theories that can be learned from data, we may also be able to develop grammatical competitors to the "vanilla" recurrent neural networks that Lau et al. (2017) present as state-of-the-art in modeling sentence acceptability judgments.

References

Adger, David. 2017. The autonomy of syntax. Unpublished manuscript. Queen Mary University, ms. http://ling.auf.net/lingbuzz/003442.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90. 119–161.

Alderete, John & Paul Tupper. 2018. Connectionist approaches to generative phonology. In S.J. Hannahs & Anna Bosch (eds.), *The Routledge Handbook of Phonological Theory*. Routledge.

Alderete, John, Paul Tupper & Stefan A. Frisch. 2013. Phonological constraint induction in a connectionist network: learning OCP-Place constraints from data. *Language Sciences* 37. 52–69.

Alhama, Raquel & Willem Zuidema. 2018. Pre-Wiring and Pre-Training: What Does a Neural Network Need to Learn Truly General Identity Rules? *Journal of Artificial Intelligence Research* 61. 927–946.

Alishahi, Afra, Marie Barking & Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. 368–378. (Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/K17-1037.

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, et al. 2016. Deep speech 2: end-to-end speech recognition in English and mandarin. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 173–182. New York, NY, USA.

[18] Pearl and Goldwater point out that the explicitly coded representational structure of Bayesian models has advantages over neural models with hidden layers in terms of interpretability, which may be especially important to linguists. The MaxEnt models discussed in section 4.2 also have interpretability benefits relative to deep neural models. Depending on the application, these benefits may outweigh advantages of neural models (which as discussed in the text, seem considerable for the learning of representations).

Anderson, James A. & Edward Rosenfeld. 2000. *Talking Nets: An Oral History of Neural Networks*. (Bradford Books). MIT Press.
Anderson, Stephen R. 1985. *Phonology in the twentieth century : theories of rules and theories of representations*. Chicago: University of Chicago Press.
Andreas, Jacob & Zoubin Ghahramani. 2013. A Generative Model of Vector Space Semantics. *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, 91–99.
Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *ICLR 2015*. https://arxiv.org/pdf/1409.0473.pdf (14 September, 2017).
Bai, S., J. Zico Kolter & V. Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv e-prints*.
Bates, E., J. Elman, M. Johnson, A. Karmiloff-Smith, D. Parisi & K. Plunkett. 1998. Innateness and emergentism. In W. Bechtel & G. Graham (eds.), *A Companion to Cognitive Science*. Basil Blackwood.
Berent, Iris. 2013. *The Phonological Mind*. Cambridge University Press.
Berent, Iris, Colin Wilson, Gary F Marcus & Douglas K Bemis. 2012. On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic inquiry* 43(1). 97–119.
Bermúdez-Otero, Ricardo. 2016. Comment on "Chomsky 1957 and the Past Tense." *Phonolist*. http://blogs.umass.edu/phonolist/2016/06/28/discussion-chomsky-1957-on-the-english-past-tense/ (24 August, 2017).
Bernardy, Jean-Philippe & Shalom Lappin. 2017. Using Deep Neural Networks to Learn Syntactic Agreement. *Linguistic Issues in Language Technology* 15(2). http://journals.linguisticsociety.org/elanguage/lilt/index.html.
Block, H. D., B. W. Knight & F. Rosenblatt. 1962. Analysis of a Four-Layer Series-Coupled Perceptron. II. *Reviews of Modern Physics* 34(1). 135–142. doi:10.1103/RevModPhys.34.135.
Block, H.D. 1962. The perceptron: a model for brain functioning. *Reviews of Modern Physics* 34. 123– 135.
Block, H.D. 1970. A review of "perceptrons: An introduction to computational geometry≒. *Information and Control* 17(5). 501–522. doi:10.1016/S0019-9958(70)90409-2.
Boeckx, Cedric. 2014. What Principles and Parameters got wrong. *Linguistic Variation in the Minimalist Framework*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198702894.003.0008. http://www.oxfordscholarship.com/10.1093/acprof:oso/9780198702894.001.0001/acprof-9780198702894-chapter-8.
Boersma, Paul. 1997. *How we learn variation, optionality, and probability*.
Boersma, Paul. 2003. Review of Tesar and Smolensky (2000), Learnability in Optimality Theory. *Phonology* 20. 436–446.
Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32. 45–86.
Boersma, Paul & Jan-Willem van Leussen. 2017. Efficient Evaluation and Learning in Multilevel Parallel Constraint Grammars. *Linguistic Inquiry* 48(3). 349–388. doi:10.1162/ling_a_00247.

Boersma, Paul & Joe Pater. 2016. Convergence properties of a gradual learner in Harmonic Grammar. In John J. McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, 389–434. Bristol, Connecticut: Equinox Publishing.

Bond, Oliver, Greville G Corbett, Marina Chumakina & Dunstan Brown. 2016. *Archi: Complexities of agreement in cross-theoretical perspective*. . Vol. 4. Oxford University Press.

Bowman, Samuel R., Gabor Angeli, Christopher Potts & Christopher D. Manning. 2015. Learning Natural Language Inference from a Large Annotated Corpus. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Stroudsburg, PA: Association for Computational Linguistics.

Bowman, Samuel R., Christopher D. Manning & Christopher Potts. 2015. Tree-structured composition in neural networks without tree-structured architectures. *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches - Volume 1583*, 37–42. Montreal, Canada: CEUR-WS.org.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford, U.K.: Blackwell.

Bybee, Joan. 1988. Morphology as lexical organization. In Michael Hammond & Michael Noonan (eds.), *Theoretical approaches to morphology*. San Diego: Academic Press.

Bybee, Joan & James L. McClelland. 2005. Alternatives to the combinatorial paradigms of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*. 381–410.

Cedergren, Henrietta J. & David Sankoff. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language* 50(2). 333–355. doi:10.2307/412441.

Chalmers, David J. 1990. Syntactic Transformations on Distributed Representations. *Connection Science* 2(1–2). 53–62. doi:10.1080/09540099008915662.

Cherry, Colin. 1957. On human communication; a review, a survey, and a criticism.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1959. A Review of BF Skinner's Verbal Behavior. *Language* 35(1). 26–58.

Chomsky, Noam. 1964. *Current Issues in Linguistic Theory*. The Hague: Mouton.

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

Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon Books.

Chomsky, Noam. 1980. On cognitive structures and their development. In M. Piatell-Palmarini (ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky*, 36–54. London: Routledge and Kegan.

Chomsky, Noam, Ángel J. Gallego & Dennis Ott. 2017. Generative Grammar and the Faculty of Language: Insights, Questions, and Challenges. *To appear in the Catalan Journal of Linguistics*. lingbuzz/003507.

Chomsky, Noam & Morris A. Halle. 1968. *The sound pattern of English*. Cambridge, Massachusetts: MIT Press.

Christiansen, Morten H. & Nick Chater (eds.). 2001. *Connectionist psycholinguistics*. Westport, Connecticut: ABLEX Publishing Corporation.

Coetzee, Andries & Joe Pater. 2011. The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.), *The Handbook of Phonological Theory*, 401–431. 2nd ed. Blackwell.

Coetzee, Andries W. 2008. Grammaticality and Ungrammaticality in Phonology. *Language* 84(2). 218–257.

Doucette, Amanda. 2017. Inherent Biases of Recurrent Neural Networks for Phonological Assimilation and Dissimilation. *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*. 35–40.

Doumas, Leonidas AA, Guillermo Puebla & Andrea E Martin. 2017. How we learn things we don't know already: A theory of learning structured representations from experience. *bioRxiv*. 198804.

Dresher, B. Elan. 1981. On the Learnability of Abstract Phonology. In C.L. Baker & John J McCarthy (eds.), *The Logical Problem of Language Acquisition*. Cambridge, MA: MIT Press.

Dresher, B. Elan. 1990. Review of Halle and Vergnaud (1987) An Essay on Stress. *Phonology* 7. 171–188.

Dresher, B. Elan. 1999. Charting the learning path: cues to parameter setting. *Linguistic Inquiry* 30(1). 27–67.

Dyer, Chris, Adhiguna Kuncoro, Miguel Ballesteros & Noah A. Smith. 2016. Recurrent Neural Network Grammars. *Proc. of NAACL*.

Edelman, Shimon. 2017. Language and other complex behaviors: Unifying characteristics, computational models, neural mechanisms. *Language Sciences* 62(Supplement C). 91–123. doi:10.1016/j.langsci.2017.04.003.

Elman, Jeffrey L. 1990. Finding Structure in Time. *Cognitive Science* 14(2). 179–211. doi:10.1207/s15516709cog1402_1.

Elman, Jeffrey L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 195–225.

Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press.

Embick, David & Alec Marantz. 2005. Cognitive neuroscience and the English past tense: comments on the paper by Ullman et al. *Brain and language* 93 2. 243–7; discussion 248-52.

Ettinger, Allyson, Sudha Rao, Hal Daumé III & Emily M. Bender. 2017. Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 1–10.

Fitz, Hartmut & Franklin Chang. 2017. Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition* 166(Supplement C). 225–250. doi:10.1016/j.cognition.2017.05.008.

Frank, Robert & Donald Mathis. 2007. Transformational networks. *Proceedings of the 3rd Workshop on Psychocomputational Models of Human Language Acquisition*. Nashville, Tennessee. http://blogs.umass.edu/brain-wars/files/2017/06/cogsci-2007.pdf.

Frank, Robert, Donald Mathis & William Badecker. 2013. The Acquisition of Anaphora by Simple Recurrent Networks. *Language Acquisition* 20(3). 181–227. doi:10.1080/10489223.2013.796950.
Gasser, Michael & Chan-Do Lee. 1990. *Networks and morphophonemic rules revisited*. (Technical Reports 307). Bloomington, Indiana: Indiana University, Computer Science Department. https://www.cs.indiana.edu/pub/techreports/TR307.pdf.
Gazdar, Gerald, Ewen Klein, Geoffrey K. Pullum & Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Cambridge, MA, USA: Harvard University Press.
Gibson, E. & K. Wexler. 1994. Triggers. *Linguistic Inquiry* 25. 407–454.
Goldberg, Yoav. 2012. Do Baboons really care about letter-pairs? Monkey-reading, predictive patterns and machine learning. https://www.cs.bgu.ac.il/~yoavg/uni/bloglike/baboons.html.
Goldberg, Yoav. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57. 345–420.
Goldberg, Yoav. 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies* 10(1). 1–309. doi:10.2200/S00762ED1V01Y201703HLT037.
Goldsmith, John. 1993. Harmonic Phonology. In John Goldsmith (ed.), *The last phonological rule: reflections on constraints and derivations*, 21–60. University of Chicago Press.
Goldwater, Sharon J. & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Erkisson & Osten Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120.
Gould, Isaac. 2015. Syntactic Learning from Ambiguous Evidence: Errors and End-States. Cambridge, MA: MIT.
Grainger, Jonathan, Stéphane Dufau, Marie Montant, Johannes C. Ziegler & Joël Fagot. 2012. Orthographic Processing in Baboons (Papio papio). *Science* 336(6078). 245. doi:10.1126/science.1218152.
Gupta, Prahlad & David Touretzky. 1994. Connectionist models and linguistic theory: investigations of stress systems in language. *Cognitive Science* 18. 1–50.
Halle, Morris & Jean-Roger Vergnaud. 1987. *An essay on stress*. Cambridge, Massachusetts: The MIT press.
Hare, Mary, David Corina & Garrison Cottrell. 1989. A Connectionist Perspective on Prosodic Structure. *Annual Meeting of the Berkeley Linguistics Society* 15(0). 114–125. doi:10.3765/bls.v15i0.1732.
Harris, Zellig. 1951. *Methods in Strcutural Linguistis*. Chicago: University of Chicago Press.
Hayes, Bruce. 1980. A metrical theory of stress rules. Massachusetts Institute of Technology.
Hayes, Bruce, Bruce Tesar & Kie Zuraw. 2013. *OTSoft*. Los Angeles: UCLA.
Hayes, Bruce & Colin Wilson. 2008. A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.
Hebb, D. O. 1949. *The organization of behavior; a neuropsychological theory, (by) D.O. Hebb. Science Editions.* New York: John Wiley and Sons.

Heinz, Jeffrey & William Idsardi. 2011. Sentence and word complexity. *Science* 333(6040). 295–297.
Heinz, Jeffrey & William Idsardi. 2013. What Complexity Differences Reveal About Domains in Language*. *Topics in Cognitive Science* 5(1). 111–131. doi:10.1111/tops.12000.
Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9(8). 1735–1780.
Hummel, John E. 2010. Symbolic Versus Associative Learning. *Cognitive Science* 34(6). 958–965. doi:10.1111/j.1551-6709.2010.01096.x.
Hunter, Tim & Chris Dyer. 2013. Distributions on Minimalist Grammar Derivations. *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*, 1–11. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-3001.
Jackendoff, Ray. 1975. Morphological and Semantic Regularities in the Lexicon. *Language* 51(3). 639–671. doi:10.2307/412891.
Jäger, Gerhard. 2007. Maximum Entropy models and Stochastic Optimality Theory. In Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson & Annie Zaenen (eds.), *Architectures, rules, and preferences: a festschrift for Joan Bresnan*, 467–479. Stanford, California: CSLI Publications.
Jäger, Gerhard & James Rogers. 2012. Formal language theory: refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1598). 1956–1970.
Jarosz, Gaja. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language* 37(3). 565–606. doi:10.1017/S0305000910000103.
Jarosz, Gaja. 2013. Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology* 30(01). 27–71.
Jarosz, Gaja. 2015. Expectation Driven Learning of Phonology. University of Massachusetts Amherst, ms. http://blogs.umass.edu/jarosz/2015/08/24/expectation-driven-learning-of-phonology/.
Jia, Robin & Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. *CoRR* abs/1707.07328. http://arxiv.org/abs/1707.07328.
Johnson, Mark. 2013a. A gentle introduction to maximum entropy, log-linear, exponential, logistic, harmonic, Boltzmann, Markov Random Fields, Conditional Random Fields, etc., models. http://web.science.mq.edu.au/~mjohnson/papers/Johnson12IntroMaxEnt.pdf.
Johnson, Mark. 2013b. *Language acquisition as statistical inference*.
Johnson, Mark, Joe Pater, Robert Staubs & Emmanuel Dupoux. 2015. Sign constraints on feature weights improve a joint model of word segmentation and phonology. *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. 2015.
Jordan, Michael. 2018. Artificial intelligence: The revolution hasn't happened yet. *Medium*. https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7 (24 May, 2018).

Kager, René. 2005. Rhythmic licensing theory: an extended typology. *Proceedings of the third international conference on phonology*, 5–31. Seoul National University.
Kaisse, Ellen M. & Patricia A. Shaw. 1985. On the Theory of Lexical Phonology. *Phonology Yearbook* 2. 1–30.
Katz, Yarden. 2012. Noam Chomsky on Where Artificial Intelligence Went Wrong. An extended conversation with the legendary linguist. *The Atlantic*. https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/.
Kenstowicz, Michael. 1994. *Phonology in generative grammar*. Cambridge, Massachusetts: Blackwell.
Keyser, Samuel J., George A. Miller & Edward Walker. 1978. *Cognitive Science, 1978. Report of The State of the Art Committee to The Advisors of The Alfred P. Sloan Foundation*. http://www.cbi.umn.edu/hostedpublications/pdf/CognitiveScience1978_OCR.pdf.
Kiparsky, Paul. 1982. Lexical phonology and morphology. *Linguistics in the morning calm*.
Kirov, Christo. 2017. Recurrent Neural Networks as a Strong Domain-General Baseline for Morpho-Phonological Learning.
Kirov, Christo & Ryan Cotterell. 2018. Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*.
Kleene, Stephen C. 1956. Representation of Events in Nerve Nets and Finite Automata. In C Shannon & J McCarthy (eds.), *Automata Studies, Annals of Math. Studies 34*.
Lachter, J. & Thomas G. Bever. 1988. The relation between linguistic structure and associative theories of language learning. *Cognition* 28. 195–247.
Lakoff, George. 1988. A Suggestion for a Linguistics with Connectionist Foundations. (Ed.) David Touretzky. *Proceedings of the 1988 Connectionist Summer School.* http://www.escholarship.org/uc/item/5df11196.
Lakoff, George. 1993. Cognitive Phonology. In John Goldsmith (ed.), *The last phonological rule: reflections on constraints and derivations*, 117–145. University of Chicago Press.
Lau, Jey Han, Alexander Clark & Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science* 41(5). 1202–1241. doi:10.1111/cogs.12414.
LeCun, Yann. 1988. A theoretical framework for back-propagation. In David Touretzky, Geoffrey Hinton & Terrance Sejnowski (eds.), *Proceedings of the 1988 connectionist models summer school*, 21–28. Pittsburgh: Morgan Kaufmann, CMU.
LeCun, Yann, Yoshua Bengio & Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553). 436–444. doi:10.1038/nature14539.
Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. In M. Ziolkowski, M. Noske & K. Deaton (eds.), *Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society*, 237–252. Chicago: Chicago Linguistic Society.

Levy, O., K. Lee, N. FitzGerald & L. Zettlemoyer. 2018. Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum. *ArXiv e-prints*.
Lewis, John D. & Jeffrey L. Elman. 2001. Learnability and the Statistical Structure of Language: Poverty of Stimulus Arguments Revisited. *PROCEEDINGS OF THE 26TH ANNUAL BOSTON UNIVERSITY CONFERENCE ON LANGUAGE DEVELOPMENT*, 359–370. Cascadilla Press.
Liberman, Mark. 2004. The curious case of quasiregularity. *Language Log*. http://itre.cis.upenn.edu/~myl/languagelog/archives/000344.html (24 August, 2017).
Lidz, Jeffrey, William Snyder & Joe Pater (eds.). 2016. *The Oxford handbook of developmental linguistics*. First edition. (Oxford Handbooks in Linguistics). Oxford: Oxford University Press.
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.
Marcus, Gary F. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.
McCarthy, J.J. 2002. *A Thematic Guide to Optimality Theory*. (Research Surveys in Linguistics). Cambridge University Press. https://books.google.ca/books?id=j_NytP1n0HoC.
McCarthy, John J, Joe Pater & Kathryn Pruitt. 2016. Cross-level interactions in Harmonic Serialism. *Harmonic Grammar and Harmonic Serialism*, 87–138. Bristol, Connecticut: Equinox Publishing.
McClelland, James L. & Karalyn Patterson. 2002. Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences* 6(11). 465–472.
McCoy, R. Thomas, Robert Frank & Tal Linzen. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *CoRR* abs/1802.09091. http://arxiv.org/abs/1802.09091.
McCulloch, Warren S. & Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5(4). 115–133. doi:10.1007/BF02478259.
Miller, George A. 2003. The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences* 7(3). 141–144.
Minsky, Marvin & Seymour Papert. 1988. *Perceptrons: an introduction to computational geometry*. Cambridge, Massachusetts: MIT Press.
Moreton, Elliott & Joe Pater. 2011. *Formally biased phonology: complexity in learning and typology*. (Ed.) K.G. Vijayakrishnan.
Moreton, Elliott, Joe Pater & Katya Pertsova. 2015. Phonological concept learning. *Cognitive Science*. 1–66. doi:10.1111/cogs.12319.
Nagy, George. 1991. Neural networks-then and now. *IEEE Transactions on Neural Networks* 2(2). 316–318.
Nazarov, Aleksei & Gaja Jarosz. 2017. Learning Parametric Stress without Domain-Specific Mechanisms. *Proceedings of the Annual Meetings on Phonology* 4(0). doi:10.3765/amp.v4i0.4010.

https://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/4010.
Nazarov, Aleksei & Joe Pater. 2017. Learning opacity in Stratal Maximum Entropy Grammar. *Phonology* 34(2). 299–324. doi:10.1017/S095267571700015X.
Neelakantan, A., L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach & J. Martens. 2015. Adding Gradient Noise Improves Learning for Very Deep Networks. *ArXiv e-prints*.
Newell, Ben R., John C. Dunn & Michael Kalish. 2011. Systems of Category Learning: Fact or Fantasy? In Brian H. Ross (ed.), *Psychology of learning and motivation: Advances in Research and Theory*, vol. 54, 167–215. Academic Press.
Norvig, Peter. n.d. On Chomsky and the Two Cultures of Statistical Learning. *norvig.com*. http://norvig.com/chomsky.html (14 September, 2017).
Novikoff, Albert B.J. 1962. On convergence proofs for perceptrons. *Proceedings of the symposium on the mathematical theory of automata*, vol. 12, 615–622. Polytechnic Institute of Brooklyn.
Olazaran, Mikel. 1993. A Sociological History of the Neural Network Controversy. *Advances in Computers* 37. 335–425.
Olazaran, Mikel. 1996. A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science* 26(3). 611–659. doi:10.1177/030631296026003005.
Palangi, H., P. Smolensky, X. He & L. Deng. 2017. Deep Learning of Grammatically-Interpretable Representations Through Question-Answering. *ArXiv e-prints*.
Pater, Joe. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39(2). 334–345.
Pater, Joe. 2016a. Universal Grammar with Weighted Constraints. In John J. McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, 1–46. Bristol, Connecticut: Equinox Publishing.
Pater, Joe. 2016b. Prince vs. Smolensky. *Brain Wars*. http://blogs.umass.edu/brain-wars/the-debates/prince-vs-smolensky/ (24 August, 2017).
Pater, Joe. 2017. Did Frank Rosenblatt invent Deep Learning in 1962? http://blogs.umass.edu/comphon/2017/06/15/did-frank-rosenblatt-invent-deep-learning-in-1962.
Pater, Joe (ed.). 2018. *Perceptrons and Syntactic Structures at 60: Collected Presentations from the 2018 Workshop*. YouTube video playlist. https://www.youtube.com/playlist?list=PL9UURLQttNX2Lfs0EoOlIa4ns0bhra8_Y.
Pater, Joe & Robert Staubs. 2013. Modeling learning trajectories with batch gradient descent. MIT, Cambridge MA. http://people.umass.edu/pater/pater-staubs-gradient-descent-2013.pdf.
Pater, Joe, Robert Staubs, Karen Jesney & Brian Smith. 2012. Learning probabilities over underlying representations. *Proceedings of the Twelfth Meeting of the ACL-SIGMORPHON: Computational Research in Phonetics, Phonology, and Morphology*, 62–71.
Pearl, Lisa & Sharon Goldwater. 2016. Statistical Learning, Inductive Bias, and Bayesian Inference in Language Acquisition. In Jeffrey L. Lidz, William Snyder & Joe Pater (eds.), *The Oxford Handbook of Developmental Linguistics*, 664–695. Oxford University Press.

http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199601264.001.0001/oxfordhb-9780199601264-e-28.
Pereira, Fernando. 2000. Formal grammar and information theory: Together again? *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY* 358. 1239–1253.
Pinker, Steven. 1984. *Language learnability and language development*. Cambridge, Massachusetts: Harvard University Press.
Pinker, Steven. 1999. *Words and Rules: the ingredients of language*. New York: William Morrow.
Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1). 73–193.
Pinker, Steven & Michael T. Ullman. 2002. The past and future of the past tense. *Trends in cognitive sciences* 6(11). 456–463.
Plunkett, Kim & Virginia Marchman. 1993. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition* 48(1). 21–69.
Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.
Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker. 2010. Harmonic Grammar with linear programming: from linear systems to linguistic typology. *Phonology* 27(1). 77–117.
Prickett, Brandon. 2017. Vanilla Sequence-to-Sequence Neural Nets cannot Model Reduplication. *UMass Open Working Papers in Linguistics*. http://scholarworks.umass.edu/ics_owplinguist/2/.
Prince, Alan & Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. Blackwell.
Prince, Alan, Bruce Tesar & Nazarré Merchant. 2015. *OTWorkplace*. New Brunswick, New Jersey: Rutgers University. https://sites.google.com/site/otworkplace/.
Pullum, Geoff & Barbara C. Scholtz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistics Review* 19. 9–50.
Pullum, Geoffrey K. 2011. On the Mathematical Foundations of Syntactic Structures. *Journal of Logic, Language and Information* 20(3). 277–296. doi:10.1007/s10849-011-9139-8.
Roeper, Thomas & Edwin Williams. 1987. *Parameter Setting*. (Studies in Theoretical Psycholinguistics 4). Springer Netherlands.
Rosenblatt, F. 1964. Analytic Techniques for the Study of Neural Nets. *IEEE Transactions on Applications and Industry* 83(74). 285–292. doi:10.1109/TAI.1964.5407758.
Rosenblatt, Frank. 1957. *The perceptron, a perceiving and recognizing automaton (Project PARA)*. Cornell Aeronautical Laboratory.
Rosenblatt, Frank. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6). 386–408.
Rosenblatt, Frank. 1962. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Washington, D.C.: Spartan Books.

Rosenblatt, Frank. 1967. Recent work on theoretical models of biological memory. In J.T. Tou (ed.), *Computer and Information Sciences II*, 33–56.
Rumelhart, D. E., G. E. Hinton & R. J. Williams. 1986. Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland & CORPORATE PDP Research Group (eds.), *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, 318–362. MIT Press.
Rumelhart, D. E. & J. L. McClelland. 1986. On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart & the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, vol. 2, 216–271. MIT Press.
Sanz, Ricardo. 2008. Top 100 most influential works in cognitive science. http://tierra.aslab.upm.es/public/index.php?option=com_content&task=view&id=141 (23 August, 2017).
Schmidhuber, Jürgen. 2015. Deep Learning in Neural Networks: An Overview. *Neural Networks* 61. 85–117. doi:10.1016/j.neunet.2014.09.003.
See, Abigail. 2017. Four deep learning trends from ACL 2017. http://www.abigailsee.com/2017/08/30/four-deep-learning-trends-from-acl-2017-part-1.html (5 October, 2017).
Seidenberg, Mark S. & David C. Plaut. 2014. Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate. *Cognitive Science* 38(6). 1190–1228. doi:10.1111/cogs.12147.
Selkirk, Elisabeth O. 1981. On The Nature of Phonological Representation. *The Cognitive Representation of Speech* 7. 379–388. doi:10.1016/S0166-4115(08)60213-7.
Shannon, CE & W Weaver. 1949. *The Mathematical Theory of Information*. University of Illinois Press.
Sharkey, Noel. 1992. *Connectionist Natural Language Processing: Readings in Connection Science*. Springer Netherlands. http://www.springer.com/us/book/9789401051606.
Smolensky, Paul. 1986. Information Processing in Dynamical Systems: Foundations of Harmony Theory. In David E. Rumelhart, James L. McClelland & CORPORATE PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, 194–281. Cambridge, MA, USA: MIT Press. http://dl.acm.org/citation.cfm?id=104279.104290.
Smolensky, Paul. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11(1). 1–23. doi:10.1017/S0140525X00052432.
Smolensky, Paul, Matthew Goldrick & Donald Mathis. 2014. Optimization and Quantization in Gradient Symbol Systems: A Framework for Integrating the Continuous and the Discrete in Cognition. *Cognitive Science* 38(6). 1102–1138. doi:10.1111/cogs.12047.
Smolensky, Paul & Geraldine Legendre. 2006. *The harmonic mind: from neural computation to optimality-theoretic grammar*. Cambridge, Massachusetts: MIT Press.
Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng & Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013*

Conference on Empirical Methods in Natural Language Processing, 1631–1642. Stroudsburg, PA: Association for Computational Linguistics.
Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever & Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15(1). 1929–1958.
Stabler, Edward P. 2013. Two Models of Minimalist, Incremental Syntactic Analysis. *Topics in Cognitive Science* 5(3). 611–633. doi:10.1111/tops.12031.
Staubs, Robert, Michael Becker, Christopher Potts, Patrick Pratt, John J. McCarthy & Joe Pater. 2010. *OT-Help*. Amherst, MA: University of Massachusetts. http://people.umass.edu/othelp/.
Staubs, Robert & Joe Pater. 2016. Learning serial constraint-based grammars. In John McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, 369–388. Bristol, Connecticut: Equinox Publishing.
Sutskever, Ilya, Oriol Vinyals & Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 3104–3112. Montreal, Canada: MIT Press.
Tesar, Bruce. 2004. Using Inconsistency Detection to Overcome Structural Ambiguity. *Linguistic Inquiry* 35(2). 219–253.
Tesar, Bruce & Paul Smolensky. 2000. *Learnability in Optimality Theory*. The MIT Press.
Touretzky, David & Deirdre Wheeler. 1991. Sequence Manipulation Using Parallel Mapping Networks. *Neural Computation* 3(1). 98–109.
Tupper, Paul F. & Bobak Shahriari. 2016. Which Learning Algorithms Can Generalize Identity-Based Rules to Novel Inputs? *CoRR* abs/1605.04002. http://arxiv.org/abs/1605.04002.
Vaux, Bert & Andrew Nevins (eds.). 2008. *Rules, Constraints, and Phonological Phenomena*. Oxford University Press.
Werbos, Paul J. 1982. Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick & F. Kozin (eds.), *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31 – September 4, 1981*, 762–770. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/BFb0006203. https://doi.org/10.1007/BFb0006203.
Wheeler, Deirdre & David Touretzky. 1993. A Connectionist Implementation of Cognitive Phonology. *The last phonological rule: reflections on constraints and derivations*, 146–172. University of Chicago Press.
Wickelgren, Wayne A. 1969. Context-sensitive coding, associative memory, and serial order in behavior. *Psychological Review* 76(1). 1–15.
Willer Gold, Jana, Boban Arsenijević, Mia Batinić, Michael Becker, Nermina Čordalija, Marijana Kresić, Nedžad Leko, et al. 2017. When linearity prevails over hierarchy in syntax. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1712729115. http://www.pnas.org/content/early/2017/12/28/1712729115.
Wilson, Colin & Gillian Gallagher. 2016. Beyond bigrams for surface-based phonotactic models: a case study of South Bolivian Quechua. https://colincwilson.github.io/papers/WilsonGallagher_sigmorphon2016.pdf.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144. http://arxiv.org/abs/1609.08144.

Xiang, Ming, Brian Dillon & Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language* 108(1). 40–55. doi:10.1016/j.bandl.2008.10.002.

Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

Yogatama, Dani, Phil Blunsom, Chris Dyer, Edward Grefenstette & Wang Ling. 2016. Learning to Compose Words into Sentences with Reinforcement Learning. *CoRR* abs/1611.09100. http://arxiv.org/abs/1611.09100.

Yu, Kristine H. 2017. Advantages of constituency: computational perspectives on Samoan word prosody. *Proceedings of Formal Grammar 2017*. http://www.krisyu.org/pages/pdfs/fg-kmyu.pdf.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28(2). 417–472.