Connectionist Models and the Nativist Debate

Christina Behme
Department of Philosophy, Dalhousie University

Abstract: *More than 50 years ago Noam Chomsky proposed that language acquisition is domain specific and depends on innate knowledge. This article introduces computational models of language acquisition challenging this proposal. Early models (e.g., Elman, 1991) showed that mechanisms not specific to language can simulate important aspects of language acquisition. More recent models using as input samples of child-directed speech have successfully simulated the performance of language-learning children in respect to both their successes and their limitations (e.g., processing of higher order recursion). Further, many current computational models directly incorporate insights from previous ones and from experiments performed with children. The computational work shows the potential to provide a model of language learning that does not depend on domain specific mechanisms and suggests that the Chomskyan dictum needs to be re-evaluated.*
Keywords: *Language acquisition. Innatism. Computational modeling. Noam Chomsky.*

Título: *Modelos Conexionistas e o Debate Nativista*
Resumo: *Há mais de 50 anos, Noam Chomsky propôs que a aquisição da linguagem fosse um domínio específico e dependente de conhecimento inato. Esse artigo apresenta modelos computacionais de aquisição de linguagem que desafiam essa proposta. Modelos iniciais (e.g. Elman, 1991) mostraram que mecanismos não específicos à linguagem são capazes de simular aspectos da aquisição de linguagem. Modelos mais recentes, utilizando como input amostras de fala dirigida a crianças, têm tido sucesso em simular a performance de crianças em processo de aprendizagem da linguagem no que diz respeito tanto aos êxitos quanto às limitações delas (e.g., processamento de recursões de níveis mais altos). Além disso, muitos modelos computacionais atuais incorporam diretamente insights de modelos anteriores e de experiências realizadas com crianças. O trabalho com computadores tem apresentado o potencial de fornecer um modelo de aprendizagem de linguagem que não depende de mecanismos de domínio específico e sugere que a máxima chomskyana tenha de ser reavaliada.*
Palavras-chave: *Aquisição de Linguagem. Inatismo. Modelamento Computacional. Noam Chomsky.*

1. Introduction

Over the last decades computational models of language acquisition have become increasingly important as a tool to challenge the Chomskyan dictum that language acquisition is domain specific and depends on innate knowledge. This dictum is based on the assumption that it is impossible to learn human language from the available input (linguistic environment) without explicit negative feedback (error correction; Gold, 1967). Therefore, Chomsky postulated that much of our linguistic knowledge is innate. The dictum can be challenged in two different ways.

First, if it can be shown that the linguistic input contains enough information so that general-purpose mechanisms can achieve human-like performance, then this supports the claim that the universal grammar (UG) postulated by Chomsky is not *necessary* for language acquisition. Second, if connectionist and/or other computational models succeed in simulating language acquisition, then this may shed light on the nature of the mechanisms involved in human language acquisition. If models also succeed in other cognitive domains, then this can lend support to the hypothesis that human language relies on domain-general mechanisms. Neither of these challenges rules out that human learners depend on UG. So the task of computational modeling is not to refute the Chomskyan dictum but to provide motivation for alternative research strategies.

In this paper I discuss computational models that simulate important aspects of language acquisition. Statistical regularities of language provide a wealth of implicit information to the young language learner. Several models rely on these statistical regularities to succeed in tasks such as speech segmentation, multiple cue integration, acquisition of complex aspects of syntax and semantics, and processing of recursive linguistic structures. Throughout I introduce criticisms of connectionist modeling and show how researchers have responded to those criticisms. I begin with an overview of the pioneering work by Jeff Elman. This work has shown the potential of connectionist modeling but did not resemble closely the conditions under which children acquire language. For this reason I also introduce several models that rely on samples of child-directed speech as input, use input from several languages, and incorporate findings from earlier modeling.

2. The beginnings of connectionist modeling and early criticism

2. 1. Elman's Early Connectionist Models

Elman (1990) has pioneered the use of connectionist models (simple recurrent networks, SRNs) for language acquisition simulation. He has argued that the use of recurrent links provides networks with a dynamic memory[1]. "In this approach, hidden unit patterns are fed back to themselves; the internal representations which develop thus reflect task demands in the context of prior internal states" (Elman, 1990, p. 179). Elman claims that connectionist networks can 'learn' to solve relatively simple problems such as the temporal version of the exclusive 'or' (XOR) function and even 'discover' syntactic and semantic features of words.

To represent temporal order in parallel- distributed language-processing models is challenging because the linguistic input is sequential. There were two previous approaches before Elman's: (i) to represent time spatially in the model or (ii) to use recurrent links so the network uses its own previous *output* in processing the current input. Elman improved on these solutions (both computationally and cognitively). He uses recurrent links from the *hidden units* (the network's *internal state*) so the current processing is influenced by the previous internal state. This is done by a *context layer* that copies the hidden unit activations (for detailed description see Elman 1990, pp. 182-186). Using this method time is represented implicitly by the effect that it has

[1] There is a long-standing philosophical debate about the justification of using intentional terminology (e.g., 'memory', 'training', 'learning', 'knowing') when referring to connectionist networks and other computational models. These concerns have been acknowledged by researchers, and currently no one takes these terms to apply in their literal meaning to computational models. For this reason I refrain from using scare quotes when referring to these terms.

on processing (the network has a 'memory'). Now the network behaviour (output) is shaped by current input and the networks previous internal states. This is done "to represent time implicitly rather than explicitly [in parallel distributed processing nets]. That is, [to] represent time by the effect it has on processing and not as an additional dimension of the input" (Elman, 1990, p. 180).

Elman showed that SRNs are capable of learning more complex language-relevant structures. For example, the networks could acquire a notion functionally equivalent to 'word' as a consequence of learning the sequential structure of letter sequences that form words and sentences but are not explicitly marked as such (Elman, 1990, pp. 191-194). And even more complex relationships are mirrored in the surface structure available to connectionist nets. The order of words in sentences reflects a number of constraints, such as syntactic structure, selective restrictions, subcategorization, and discourse considerations (Elman, 1990, p. 194). The networks were able to 'learn' word order and simple sentence structure and to categorize syntactic information (noun vs. verb categories) and semantic information (foods, animals etc.), based on cues available in the surface forms. This indicates that the information about alleged 'deep' structure is implicit in the surface structure of spoken language and some aspects of language can be learned based on surface structure alone (Elman, 1990, pp. 194 -203).

2.2. Criticism of Early Connectionist Simulations

Impressive as Elman's simulations might be, one needs to be clear about what they show and what they do not show. This is important because much criticism has been directed at claims that have not been made by Elman (or other connectionists).

One of the most severe criticisms is that connectionists are empiricists who advocate that the mind is originally a blank slate. Hence, connectionist work can not be relevant to human cognition. This view has been expressed repeatedly by Chomsky: "empiricism insists that the brain is a *tabula rasa*, empty, unstructured, uniform at least as far as cognitive structure is concerned" (Chomsky, 1977, p. 2) and there are "empiricist currents, that would have us believe that the human mind is empty, a *tabula rasa*" (Chomsky, 1980a, p. 270). More recently the same view was expressed by James McGilvray[2]: "[connectionists'] claim that the mind is made up of 'neural nets' is innocuous; it is their claim about the initial state of the net (undifferentiated, approximating Locke's blank slate) and their view about how this net gets its 'content'... that place them firmly in the empiricist camp" (McGilvray, 2009, p. 110).

McGilvray seems to assume that untrained networks are blank slates, because connectionist learning starts with random weights. Inputs produce random activations and the output errors are used to adjust the connection weights. A network at time 0 would be a blank slate. However, the assumption that entirely unconstrained learning can produce any interesting results is incorrect. Some connectionists may have believed that in the early days of connectionism (Elman, p.c.), but it was quickly discovered that some constraints need to be built in the initial network (see below). So even at time 0 it is not a blank slate.

My extensive review of recent connectionist literature finds no evidence for the blank-slate position. It reveals, instead, that several researchers have explicitly or implicitly rejected completely unconstrained 'blank slate' views of language acquisition (e.g., Hare & Elman, 1995; Elman et al., 1996; Redington & Chater, 1998; MacWhinney, 2000; McDermott, 2001; Solan et al, 2005; Edelman & Waterfall, 2007; Chater & Christiansen, 2009). Explicit rejections of the blank slate view could indicate that early connectionist work might have had this flaw but that it has been corrected by now. Several points are important when dealing with this possibility.

[2] Claims to the contrary not withstanding (e.g., that he has "repeatedly, consistently, and clearly insisted that all rational approaches to the problem of learning, including 'associationism' and many others ... attribute innate structure to the organism" (Chomsky, 1983, p. 310)), Chomsky still implicitly holds that connectionists are committed to the blank slate view, because he did not object to this remark by McGilvray in the introduction to *Cartesian Linguistics* which Chomsky read and commented on "early in 2008" (McGilvray, 2009, p. 6).

First, there is an ontological point. While some connectionist researchers may initially have claimed that networks were basically blank slates, this was a misunderstanding on their part. In fact, no network is ever a blank slate. There are always built-in constraints, although they may not always be recognized. These constraints can take the form of 'maximize discriminability of input strings', 'minimize processing time', etc. Constraints also can be in the learning algorithm itself. Further, the structure of the network's architecture also provides a very real and powerful constraint on what can be learned (Elman, p.c.). Thus, even though connectionists and critics alike may have *believed* at one point that nets were *tabulae rasae*, McGilvray's claim that the initial state of connectionist nets 'approximates Locke's blank slate' is incorrect.

Second, in addition to the constraints that were 'built into' the connectionist networks, the models acquired additional structure through their interaction with the language input. Elman discussed this issue already in 1990. His claim was not that a completely unstructured neural net could acquire any language related structure. Instead, he claimed that *some* structure that was not initially in the network could be acquired through repeated exposure to language-like input. For example, Elman's work showed that nouns and verbs produce different activation patterns (in trained SRNs). These aspects of language structure were not initially present in the networks, but learned from exposure to the input. If one takes Chomsky's surface/deep structure distinction seriously, accounting for the fact that the interaction with the input changes the structure of the networks is important. Yet, Elman was able to show that a similar structuring arises over time when connectionist networks are exposed to language input:

> The representations need not be 'flat,' atomistic, or unstructured. The sentence task demonstrated that sequential inputs may give rise to internal representations which are hierarchical in nature. The *hierarchy is implicit in the similarity structure of the hidden unit activations* and does

> not require an a priori architectural commitment to the depth or form of the hierarchy. Importantly, distributed representations make available a space which can be richly structured. (Elman, 1990, p.208, emphasis added)

Elman does not claim that nets are initially completely unstructured blank slates but that they do not need to have one specific fixed (innate) structure in order to solve the sentence task. He suggests the task does not require such narrow structuring because in the case of language much of the structure is contained in the input: "What is exciting about the present results is that they suggest that the inductive power of the PDP approach can be used to *discover structure* and representations in tasks which unfold over time" (Elman, 1990, p. 209, emphasis added). For Elman the connectionist models can help uncovering how much of the structure contained in language output can be acquired from the structure of the input. But this does not entail that just *any* network could succeed in this task. Other researchers have stressed the same point:

> [models] provide insight into which aspects of network performance are due to architectural biases and which arise due to learning. A network always has some bias with respect to a particular task, and this bias is dependent on a number of factors, such as overall network configuration, the nature of the activation function(s), the properties of the input/output representations, the initial weight setting, etc. (Christiansen & Chater, 1999, p. 195)

Thus, the simulations will also help to discover how much structure needs to be 'built in' to the networks and whether or not this structure needs to be task specific. This leads to a second common criticism of connectionist models. It is often alleged that general-purpose learning mechanisms are, in principle, not able to solve the language-learning task (e.g., Chomsky, 1959, 1966, 1975a, 1986a, 2005, 2012; Marcus, 1993; Smith, 1999; Crain & Pietroski, 2002; McGilvray, 2005, 2009).

This criticism was initially leveled against behaviourism. But over the years it has become a criticism of any language-acquisition account that does not posit innate domain-specific knowledge and/or mechanisms. Consider:

> ...people argue that environmental factors are critical but without offering any account of the facts in question in terms of such alleged factors. And as long as they don't produce any moderately plausible account in terms of presumed environmental factors, all I can say is that they're not holding my attention. It is not very interesting if somebody claims that something is the result of the environment or an act of God or electrical storms in the vicinity, or whatever, if they don't provide some explanatory scheme that can at least be investigated. (Chomsky, 1993, p. 4)

Examining Elman's work reveals that he did not commit the sins Chomsky (1993) alleges. Elman (1990, 1993, 1999) has attempted to show that the language input (=environment) contains information that is relevant and important for language acquisition. The input does to some degree determine the output. However, Elman also has shown that the relationship between input and output is not merely one of stimulus-response. If this were the case, the connectionist nets would fail in any tasks that require dealing with previously unencountered examples. Elman claims that some of the information that allows the nets to deal with new examples is contained in the input. However, he never claimed that *all* the relevant information is 'in the environment'. Instead, he specifically acknowledges that the information contained in the input alone is insufficient:

> While it is undoubtedly true that the surface order of words does not provide the most insightful basis for generalizations about word order, it is also true that from the point of view of the listener, the surface order is the only visible (or audible) part. Whatever the abstract underlying structure be, it is cued by the surface forms, and therefore, that structure is implicit in them. (Elman, 1990, p. 195)

Here Elman suggests that it might be possible to use regularities "on the surface" (the language available as input) to uncover regularities in the ‘abstract underlying structure’. It is of course possible that this approach will turn out to be wrong. However, it is not the case that Elman does not “produce any moderately plausible account in terms of presumed environmental factors” (Chomsky, 1993, p. 14). Quite to the contrary, he offers a detailed account of the input (= presumed environmental factor) he uses for his simulations (Elman, 1990, pp. 187-188, 193, 195-196, 200). Furthermore, it is not true that he appeals vaguely “to ‘similarity’ or ‘generalization’ without characterizing precisely the ways in which a new sentence is ‘similar’ to familiar examples or ‘generalized’ from them” (Chomsky, 1971, p.48). Quite to the contrary, Elman carefully specifies in which ways the novel sentences are similar to those the nets encountered during training (Elman, 1990, pp. 195 - 197). He outlines details regarding the experimental procedure, the input, the expected output and the actual performance of the nets.

Elman continued his work on SRNs and explored, for example, the learning of sentences with embedded clauses (Elman, 1991), and complex embedded structures (Elman, 1993). The latter work showed that tasks, traditionally thought to require an explicitly recursive computational structure, could be solved by the simple network architecture of an SRN. Here Elman also attempted to implement the insight that in humans “learning and development interact in an important and non-obvious way. Maturational changes may provide the enabling conditions which allow learning to be most effective” (Elman, 1993, p. 72). He demonstrated that in some circumstances, SRNs that are trained to represent part/whole relationships and embedded clauses of complex sentences “work best when they are forced to ‘start small’ and to undergo a developmental change which resembles the increase in working memory which also occurs over time in children” (Ibid.). In other cases, ‘learning’ can only occur when the entire data set is available to a network (e.g., Harris, 1991). Elman claims that “the deeper principles which underlie learning in the general class of connectionist systems which rely on error-driven gradient descent techniques... interact with characteristics of human development in a beneficial manner” (Elman, 1993, p. 72). Given that infants also start with limited memory capacity and only pay attention to a small segment to the linguistic input they receive, connectionist networks simulate one important aspect of human learning in general and language acquisition in particular.

Elman’s approach has been critiqued by Rohde and Plaut (2003), who suggest that artificial languages that only contain the relevant syntactic information are not a good representation of human languages. Human subjects rely to a considerable degree on semantic information when processing sentences that contain long-distance dependencies and it is questionable that SRNs that are deprived of access to semantic information perform a task that is sufficiently similar to the task faced by human children. For this reason Rohde and Plaut performed experiments using an artificial language that provided syntactic as well as semantic information[3] (for details see Rohde & Plaut, 2003, pp. 2-4). Unlike Elman they found that networks that were exposed to complex constructions throughout training outperformed those that ‘started small’. “Under no condition did the simple training regimen outperform the complex training” (Ibid., p. 5). These authors believe “that recurrent connectionist networks already have an inherent tendency to extract simple regularities first” (Ibid., p. 18). The network first learns short-range dependencies and considers long-range constraints as noise. Once the short-range dependencies are learned, the network can use the available information to learn long-distance dependencies. This is an ‘innate’ constraint, but it is not domain specific. Similarly, in children the specific stages of language acquisition could be caused by a cognitive “system that is unorganized and

[3] One of the models used stochastic units and a single layer of weights and learned to map from the semantic features of the three or four main constituents of the sentence to the semantic representations for the fillers of up to four thematic roles: agent, patient, instrument, and modifier (p. 7).

inexperienced but possesses great flexibility and potential for future adaptation, growth and specialization" (Ibid., p. 21).

The SRN simulations discussed so far focus on the acquisition of syntactic structure, which is just a small part of the overall language-learning process. The promising results do not imply that all aspects of language acquisition can be modeled by connectionist nets. But they challenge the claim that "there is no reason to take empiricist speculations at all seriously ... [because] the apparent aim is not to explain facts of human language and concepts and their growth" (McGilvray, 2009, p. 23). If it can be shown that the 'learning' in connectionist nets resembles that of human children in important aspects, then this work should be taken seriously. The findings of connectionists can help to determine the path of further research. For example, this research can help to determine whether or not we need to postulate complex internal representations in order to account for language acquisition and processing. Using a simulation that did not contain such representations was an exploratory step with not necessarily expected results: "The approach described here employs a simple architecture, but is surprisingly powerful" (Elman, 1990, p. 207).

Elman's models allow us to draw some conclusions regarding the ability of simple mechanisms to acquire language relevant 'knowledge'. However, they were neither replicating what children do when they acquire language nor based on the input children would typically receive. In the next section I discuss models that use child-directed language as input and better address the challenge that modeling has nothing to do with the conditions under which children acquire language. I focus on precise models "that can be tested, refined or rejected on the basis of publicly available data and/or replicable experiments" (MacWhinney, 2010, p. 477).

3. Recent computational models of language-acquisition

Over the last 20 years researchers have refined computational models and many of these models rely on "corpora of spontaneous adult–child interactions made available through the Child Language Data Exchange System (CHILDES)" (MacWhinney, 2010, p. 477). These CHILDES corpora consist of recorded samplings of adult speech that serves as input to language-learning children. MacWhinney suggests that "the job of the computational modeler is to determine a set of algorithms that can take the child-directed speech (CDS) as input and produce the learner's output (LO) at successive developmental levels" (MacWhinney, 2010, p. 477). This means the input for the models resembles closely the input children receive, and the models are informed by the developmental stages that are typical for children. Children have to acquire a complex set of skills long before they are able to comprehend and produce complex grammatical structures. Recent computational models attempt to simulate several of the important steps that children take on the road to language. It would lead too far afield to discuss all relevant models here, so I focus on some of the important milestones.

3.1. Models of Speech Segmentation

Recently researchers have attempted to use computational models to simulate speech segmentation acquisition. Infants need to learn to segment the continuous stream of language input into individual words. They master this skill in the course of several months. Potentially there are many ways to achieve speech segmentation. But only modeling some of them will give us a better understanding of how children might accomplish this task. Researchers are aware of this and have tested a variety of them by now: "Previous developmental models of speech segmentation differ substantially across a number of parameters, including whether the model builds a lexicon, segments words by clustering smaller units or breaking down larger units, or incorporates external constraints on performance" (Monaghan & Christiansen, 2010, p. 546). The performance of these models on essential criteria differs. For example, some models achieve a high degree of precision (correct identification of words from the input) but rely on mechanisms that are not psycholinguistically plausible (e.g., access to the complete input, no memory limitations, optimal learning). However, even psycholinguistically implausible models can provide us with a better understanding about the information content of the input. If it turns out that the information present in the input alone is sufficient for speech segmentation, then this task does not necessarily depend on innate domain specific 'knowledge'.

Monaghan and Christiansen assume that the information contained in the input is helpful in the acquisition of speech segmentation. However, they are not only looking for a mechanism that can extract the relevant information from the input but are also attempting to provide a model (PUDDLE) that closely resembles how children accomplish this task. As input for this model they use speech spoken by adults in the presence of children aged 2 years, 6 month or younger (six English CDS corpora from the CHILDES database) (MacWhinney, 2000; for complete details see Monaghan & Christiansen, 2010, pp. 552-554). This input for the model is similar to input received by children who learn language.

PUDDLE is similar to young children in several aspects. Like children, the model builds its lexicon incrementally from the input. This 'strategy' does not require that the model makes multiple, simultaneous decisions about the match between a given utterance and the acquired lexicon. Just like young children, the model is initially unable to perform complex cognitive tasks simultaneously. PUDDLE simulates how children can take advantage of features that are readily accessible in CDS and can accommodate learning. The model performs like a child because "the memory resources and computational requirements are minimal" (Monaghan & Christiansen, 2010, p. 248)

Monaghan and Christiansen were especially interested in two of the readily accessible features of the input: (i) utterance boundaries and (ii) the interspersal of high frequency words in speech (Ibid.). As the results show, these two cues in combination can go a long way towards correct speech segmentation. This is because CDS contains a relatively high percentage of single word utterances (26% in the CHILDES corpus), and some of these words occur frequently (e.g., the child's name). Treating utterances as words and recording the frequency of words allows PUDDLE to build up a lexicon incrementally, and items from the lexicon are in turn used to determine 'new' words (e.g., parts of utterances that precede or follow an item already in the lexicon). It was possible to show that a small set of frequently occurring words can help in "carving up the rest of the speech stream into its constituent words" (see Monaghan & Christiansen, 2010, p. 250, for full details). The performance of this model, which relies on a very simple algorithm for discovering words, was impressive. Depending on the corpus used, the recall was 70-79%, meaning that PUDDLE identified between 70-79% of the words that were contained in the input. Precision was 70-76%, meaning that 70-76% of the words identified by PUDDLE were indeed words from the input. This may sound like substantially less recall and precision than children might achieve. However, we need to keep in mind that the model was only exposed to 10.000 utterances in total and that it did not have access to numerous other cues that are used by children (e.g., acoustic, phonological and prosodic cues; for overview see Monagahan & Christiansen, 2008). These results suggest that all the information contained in the input might be sufficiently rich for speech segmentation and that this information might be extracted with relatively simple mechanisms.

One might object that there is not enough acoustic variability in the input used for PUDDLE. Children have to succeed with the word-segmentation task after exposure to input that varies between speakers, and even between different utterances of the same speaker (Newman et al., 2006; Newman, 2005; Brent & Siskind, 2001; for proposed solution see Christiansen & Allen, 1997). While competent speakers can rely on context and other cues to disambiguate and/or recognize unclearly or mispronounced words, infants lack this knowledge. For modeling, the CHILDES data are transcribed, and in the process incorrect utterances, mispronounced words, etc. are eliminated. It has been argued that models relying on this 'sanitized' input do not provide a realistic simulation of language learning (McGilvray, 2009; Chomsky, 2007; Smith 1999). To address this challenge Rytting et al. (2010) developed a method to test whether models can deal with probabilistic, speech-derived input which "is more realistic than the types of transcripts usually used to train and test computational models of word segmentation" (Ibid., p. 514). They used as input "recordings of four mothers... directed at infants age 0;9 to 0;10.26" (Ibid., p. 525). From this input they removed 'problematic utterances' such as "whispered or sung speech; unintelligible, untranscribed or partial words; word play or pet names... [leaving] 13,443 utterances for the four mothers" (Ibid.), but left utterances that were clearly audible yet grammatically incorrect or incomplete. They found that the performance of SRNs and models used by other researchers (e.g., Christiansen et al. 1998) "is robust for data with subsegmental variation when this variation is carefully controlled" (Ibid., p. 530), but that an increase in variation leads to a significant degradation of the performance.

Rytting et al. hypothesized that increased variability in the input "compromises the reliability of the segmental cues, such that it is no longer possible to find word boundaries using these cues alone" (Ibid., p. 531). The authors then tested the impact of additional cues (e.g., dictionary-derived stress cues and hyperarticulation of word-initial syllable cues) and found that when the models are "faced with highly variable, potentially ambiguous input, multiple probabilistic cues still outperform each cue separately" (Ibid., p. 536). These findings indicate that natural language contains an abundance of cues for word-segmentation and that the combination of several of these cues makes the segmentation task easier even under conditions that are less than ideal. Furthermore, infants also might be "able to detect regions of clear speech, and treat the beginnings of such regions as likely word boundaries" (Ibid., p. 540).

3.2. Multiple Cue Integration Models

Research with children has shown that language contains multiple statistical cues and that children are able to access this information (e.g., Monaghan & Christiansen, 2008). Having access to more than one source of information can simplify the language-learning task because children learn over time to integrate the information from several cues into a coherent whole. Given that multiple cue integration assists children in language learning, it is desirable to simulate this process in computational modeling. One might predict that models that can access the multiple sources of statistical information that are contained in the language input will out-compete models that rely only on one source of information. Recently many researchers have begun to test whether it is possible to simulate the effects of multiple cue integration with computational models (for overview see Monaghan & Christiansen, 2008). The challenge is to design models that can access simultaneously several cues and combine the information to assist word-segmentation. Blanchard et al. (2010) propose that infants can learn individual words based on frequent occurrence (e.g., their name, 'mom', frequent function words) and/or language specific phonotactic constraints (stress patterns, allophonetic variation, etc. as discussed in chapter 4). According to Blanchard et al., frequently occurring words form the first tiny lexicon, which allows the learner to infer some phonotactic constraints. This information in turn can help to recognize additional words. This combination of these two cues solves one important problem that beginning language learners face: how can they know which phonotactic constraints apply before they know words and vice versa. Thus, "knowledge of familiar words, combined with increasingly refined phonotactic constraints, support and reinforce each other in speech segmentation" (Blanchard et al., 2010, p. 491).

Blanchard et al.'s model PHOCUS relies on very basic assumptions about language learning. Beginning with an empty lexicon, it incrementally adds items to the lexicon, based on phonemes that occur together (probabilistic and phonotactic cues). Phonemes that occur within frequent words have high transitional probabilities while phonemes that cross word boundaries have low transitional probabilities (e.g. Saffran et al., 1996). In addition to these transitional probabilities, which are helpful to detect word boundaries, Blanchard et al.'s model could exploit phonotactic cues. Specifically, when the model encountered an unfamiliar word, it could rely on two kinds of phonotactic cues (phoneme combinations and occurrence of at least one syllabic sound per word; for details see Ibid., pp. 496 - 501).

Blanchard et al. could show that the combination of these two simple cues allowed a performance of 76-81% precision/recall scores for an English test corpus. Unexpectedly, the same model performed substantially worse (19 - 47% precision/recall scores) on a Sesotho corpus (Ibid,. p. 503). This result "highlights the importance of testing acquisition models on data from a variety of languages because the results can be so different from what is obtained with English corpora" (Ibid., p.505). The authors explain this difference in performance with the fact that the most frequent word in the Sesotho sample is monosyllabic. This results in a very high percentage of over-segmentation errors from which the model cannot recover. Obviously, children learning Sesotho are able to master the word segmentation task. This indicates that they cannot rely on the same cues as the model used by Blanchard et al.. A model that incorporates more cues than the Blanchard et al. model simultaneously might succeed in the Sesotho word segmentation task.

The finding that a model that successfully simulates one aspect of language acquisition in one language may perform poorly on the same aspect when exposed to a different language is important because any child can learn any human language (Chomsky, 2005, 2007, 2012; McGilvray 2006; Smith 1999). Children acquiring different languages have to rely on different cues because each language has its own hierarchy of cues due to lexical, phonological specificities. Thus, computational modeling has shown that the hypothesis that language learning follows essentially the same steps regardless of the particular language learned needs to be re-evaluated. We cannot rely on *a priori* assumptions about the uniform nature of language learning. Instead we need to develop testable hypotheses and models and adjust these in light of the obtained results.

3. 4. Modeling of Complex Aspects of Syntax and Semantics

It has been shown early that some aspects of syntax acquisition can be modeled. One of the earliest models (Rumelhart & McClelland, 1986) simulated the acquisition of the English past-tense. It showed that a two-layered feed-forward neural network architecture could learn mappings between phonological representations of stems and corresponding past tense forms of English verbs and simulate frequency effects and the U-shaped learning curve for the acquisition of irregular verbs. These early models were criticized for their psycholinguistic implausibility (e.g., Pinker & Prince, 1988) and later models addressed this criticism. One of these later models (Plunkett & Juola, 1999) was using a single-system connectionist network to produce the plurals and past tense forms of a monosyllabic English nouns and verbs. This model mimics important features of the acquisition of English noun and verb morphology in young children (e.g., initial error-free period of performance followed by a period of intermittent over-regularization of irregular nouns and verbs) and acquires nouns and verbs in a similar manner as young children do. For example, the network exhibits a general advantage in acquiring noun morphology before verb morphology and the network model predicts the same developmental shift in the relative ease of learning irregular nouns and verbs as observed in human children.

More recently it has also been shown that combining models can increase the generality of the models across inflection types, grammatical classes, and across languages. Karaminis and Thomas (2010) combined elements of previous connectionist models of morphology to implement a generalized inflectional system. Their “Multiple Inflectional Generator” (MIG) considered three grammatical classes (nouns, verbs, and adjectives) and multiple inflections for each grammatical class (e.g., nouns: base forms, plurals, and possessives) (Karaminis & Thomas, 2010, p.732). Their preliminary results showed that models MIG can reproduce error patterns and accuracy levels of inflection acquisition:

> In both English and Modern Greek, an Optional Infinitive stage was observed, even though the character of that stage is different in each language (unmarked stems vs. 3rd person singular). Generalization rates of the past tense rule were high for novel stems, even for phonotactically illegal stems. MIG captured the order of emergence of different inflection types for different grammatical classes. And it was able to capture developmental patterns for two languages of different morphological complexity. (Karaminis & Thomas, 2010, p.734)

Critics of connectionist work might argue that, more complicated tasks could remain forever beyond the capability of computational modeling. A complete discussion of this question would require addressing the problem of artificial intelligence and is beyond the scope of this paper. Instead, I will highlight now some recent work involving the acquisition of complex aspects of syntax and semantics. I begin with two models that simulate peculiarities of French spelling and English grammar respectively. In both cases performance that seems to indicate explicit knowledge of abstract rules was achieved by mechanisms that rely only on statistical features of the input.

Computer simulations of certain aspects of language acquisition are most useful when they model closely the relevant behaviour of children. For this reason Pacton et al. (2001) tested first whether children from kindergarten to grade five use statistical cues to track orthographic language regularities. They found that kindergarten children were able to use statistical regularities in the input to judge the ‘word-likeness’ of nonsense letter strings. Consistent with the statistical information available in the input, children judged letter strings as less word-like when they began with a double vowel than with a double consonant. Further, children from grade 2 onward made this distinction even for consonants that are never doubled. Pacton et al. duplicated these results with connectionist networks (SRNs). Like the children, the SRNs became sensitive to the frequency and placement of doubled letters. The SRNs lack a mechanism for rule-based abstraction but still successfully simulate the performance of children. This led Pacton et al. to conclude that domain specific mechanisms for rule-based abstraction are unnecessary to account for this aspect of language performance.

Another aspect of language acquisition that seems to push the limits of computational modeling is the acquisition of complex verb-argument structure, such as the prepositional dative (PD) vs. direct object dative (DOD). Many English verbs (e.g., tell, give, throw, bring) can occur in both constructions. For example 'to give' occurs in "Stefan gives a book to Katrina" (PD) and in "Stefan gives Katrina a book" (DOD). However, some verbs (e.g. confess, take, say, send) occur only in PD. Thus, when encountering a new verb, children cannot reliably generalize from previous examples because they do not know in which of the two groups the new verb belongs. Yet, seemingly, children are able to use novel verbs correctly, which has lead several researchers (e.g., Pinker, 1989; Gordon, 1990; Smith, 1999; McGilvray 2006) to conclude that some form of innate knowledge is required to explain this observed performance. Thus, the question arises whether computational models can achieve a child-like performance when acquiring PD and DOD constructions. Perfors et al. (2010) present a domain-general hierarchical Bayesian model for the acquisition of PD and DOD construction. These researchers take as point of departure the work of Wonnacott et al. (2008). Wonnacott et al. found that human listeners are sensitive to distributional cues in language input and can use these cues to make productive generalizations. The computational model of Perfors et al. was informed by the knowledge that had been gathered from work with human children. They developed a computational model that explains "the acquisition of verb constructions as a rational statistical inference" (Perfors et al., 2010, p. 609).

Essentially this model can take advantage of positive and implicit negative evidence that is provided in the input. It keeps track of whether or not a given verb occurs in PD and DOD constructions. Assuming that the permissible usage is fixed, the model can use new data to make increasingly better predictions. “Each time a verb is encountered in one of the two competing structures, it is *not* encountered in the other, and this provides cumulative evidence against a grammar that allows this usage” (p. 630, original emphasis). This results in a learning outcome closely resembling that of human children: performance is very good for frequently occurring verbs but poor for verbs that are rarely encountered. Even though the model is capable of learning the distinction between alternating and non-alternating verb classes on the basis of syntactic input alone, the authors do not suggest that children exclude semantic information when learning this distinction. But the fact that it is possible for a relatively simple model to simulate this aspect of language acquisition suggests that the input contains an abundance of statistical information that can be used for inferences even before the learner has access to semantic information.

Reali and Christiansen (2005) suggest that the language input contains rich indirect statistical information that can be accessed by the language learner. Using the example of auxiliary (AUX) fronting in complex polar interrogatives they show that simple learning devices, such as neural networks, are capable of exploiting such statistical cues. This is an important finding because the problem of polar interrogatives has played an important role in language acquisition debates since the 1960s. In English declarative sentences are turned into questions by fronting the correct auxiliary. Chomsky (1965) used the following example to illustrate the problem. The declarative sentence ‘*The man who is hungry is ordering dinner.’* is grammatically turned into a question in (1a) but ungrammatically in (1b):

(1a) *Is the man who is hungry ordering dinner?*
(1b) **Is the man who hungry is ordering dinner?*

Because this fronting is not based on a structure-independent rule that could be readily learned from the available input but on a structure-dependent rule (Chomsky, 1980), it has been argued that the knowledge allowing children to produce correct auxiliary questions must be innate (e.g., Chomsky 1965, 1980; Crain 1991; Lightfoot, 1991; Smith, 1999; Crain & Pietroski, 2001; Crain & Pietroski 2002; Legate & Yang 2002; McGilvray, 2006).

Reali and Christiansen (2005) show a possible alternative to this suggestion. They trained simple statistical models based on pairs (bigrams) and triples (trigrams) of words of child-directed speech. Then they tested the models on sentences that consisted of correct polar interrogatives (e.g., *Is the man who is hungry ordering dinner?*) and incorrect ones (e.g., *Is the man who hungry is ordering dinner?*) that had not been present in the training corpus (Reali & Christiansen, 2005, p. 1010). They found that the models classified correctly 96 out of 100 grammatical test sentences and concluded that these "results indicate that it is possible to distinguish between grammatical and ungrammatical AUX questions based on the indirect statistical information in a noisy child-directed speech corpus containing no explicit examples of such constructions" (Ibid., p. 1014). Furthermore, their models were also able to simulate the production of grammatical AUX questions. This performance is based on frequency patterns in the input. Sentence chunks that are frequently encountered create a bias towards grammatical question-production even in the absence of direct positive evidence. Assuming that the models do not have innate knowledge of grammar, it seems to follow that the statistical information that is explicitly and implicitly available in the input can be used to produce grammatical AUX questions.

Many other researchers have reported similar results (e.g., Redington et al., 1998; Ellefson & Christiansen, 2000; Mintz, 2002; Perruchet & Vinter, 2002; Christiansen & Kirby, 2003). Again, it is important to establish that modeling closely mirrors abilities and strategies used by human learners. Mintz (2002) showed that adults who learned an artificial language naturally formed abstract grammatical categories solely based on distributional patterns of the input data. This could be evidence for rapidly engaged distributional mechanisms that also play a role in the early stages of language acquisition when the learner lacks access to other information (semantic and syntax). Mintz claims that his experiment "shows evidence of categorization mechanisms that function from distributional cues alone" (Mintz, 2002, p. 684).

It has been shown that computational models can replicate other important aspects of human performance. Lewis and Elman (2001) trained simple recurrent networks on data from an artificial grammar. This generated questions of the form "AUX NP ADJ?" and sequences of the form "Ai NP Bi". During training the SRNs encountered no relevant examples of polar interrogatives. In this experiment it has been shown that the SRNs were better at making predictions for multi-clause questions involving correct auxiliary fronting than for those involving incorrect auxiliary fronting.

Christiansen and Kirby (2003) demonstrate that a general model of sequential learning that relies on the statistical properties of human languages can account for many aspects of language learning. Similar to the experiments on statistical learning discussed above, artificial-language-learning experiments showed that human subjects and SRNs that were trained on ungrammatical artificial languages made significantly more errors when predicting the next word of a string than subjects and SRNs that were trained on grammatical artificial languages. Languages are considered grammatical when they contain (at least one of the) universal properties of natural languages (e.g., branching direction, subjacency) and ungrammatical when they lack these properties. The authors suggest that the close performance similarities of human subjects and SRNs could indicate that both rely on similar learning mechanisms.

Ellefson and Christiansen (2000) demonstrated that SRNs were significantly better at predicting the correct sequence of elements in a string of a 'natural language' than of an 'unnatural language'. The 'natural language' contained subjacency constraints, and the 'unnatural language' lacked these constraints. For example, SRNs trained on languages containing the 'natural' patterns exemplified in sentences (2) and (5) did significantly better than those trained on languages allowing the 'unnatural' patterns exemplified in (3) and (6):

(1) Sara heard (the) news that everybody likes cats.
(2) What (did) Sara hear that everybody likes?
(3) *What (did) Sara hear (the) news that everybody likes?
(4) Sara asked why everyone likes cats.
(5) Who (did) Sara ask why everyone likes cats?
(6) *What (did) Sara ask why everyone likes?
(Ellefson & Christiansen, 2000, p. 349f).

Ellefson and Christiansen were able to show that SRNs trained on the same input data are sensitive to the statistical properties of the input and perform significantly better on grammatical than on ungrammatical test sentences.

Other researchers obtained similar results. For example, Perruchet and Vinter (2002) defend a plausible model (PARSER) that extracts 1-5 syllable words from language input based on the information that is contained in small chunks of the l input. They demonstrate that complex material can be processed as a succession of chunks that are comprised of a small number of primitives[4]. According to these authors, associative learning mechanisms can fully account for this aspect of language learning. When SRNs and other computational models are able to acquire statistical knowledge of the input based on positive examples alone, then it seems to be at least imaginable that children can pick up this information as well. Whether or not children rely on the same mechanisms as SRNs remains a point of debate (for some critical suggestions see Marcus, 1999). But the existence of these mechanisms, again, casts some doubt on the *necessity* of UG.

[4] During the performance of the segmentation task repeated perceptual chunks evolve into processing primitives which in turn determine the way further material is perceived.

We have seen that not only word-segmentation but also the acquisition of complex grammatical forms could be based on statistical properties of the input. Statistical learning occurs in several species. It has been confirmed in non-human animals (Hauser et al., 2001; Terrace, 2001), and there is evidence that it might have been recruited to support language learning in human infants (Saffran et al., 1996; Fiser & Aslin, 2002; Maye et al., 2002). Whether or not statistical learning mechanisms can account for all aspects of language acquisition is a matter of ongoing debate. While this debate is far from over, it has become clear that proponents of the LAD need to rule out the possibility that data-driven general-purpose learning mechanisms such as statistical learning can account for the acquisition of human language.

3.5. Recursion

In the final section of this paper, I discuss some attempts to model one of the often-cited ‘hallmarks’ of human language: recursion. Recursion has played a central role in Chomsky’s arguments for the uniqueness of language (Chomsky, 1966, 1975, 1980, 1986, 2012; Hauser, Chomsky, & Fitch, 2002; Fitch, Chomsky & Hauser, 2005). It allows for unbounded linguistic creativity and remains at the core of the Minimalist Program (Chomsky, 1995). Thus, the possibility that this unique feature of language can be simulated by computational models casts some doubt on the proposal that recursion is necessarily an “innate property of grammar or an a priori computational property of the neural systems subserving language” (Christiansen & MacDonald, 2009, p.127).

There are different types of recursive constructions as well as several levels of complexity within these types. Left- and right-branching recursion (LBR, RBR) is fairly common in many languages. One complex example of RBR in English is: “This is the dog, that chased the cat, that killed the rat, that ate the cheese, that lay in the house that Jack built” (Sampson, 2001, p.133). Even though this sentence involves 4 levels of RBR it can be processed and understood by an average native speaker, and will generally be judged as grammatical.

Center embedded recursion (CER) on the other hand is more difficult to process, as this example shows: “When the pain, that nobody, who has not experienced it can imagine, finally arrives, they can be taken aback by its severity” (Ibid., p. 20). This sentence only involves 2 levels of CER, yet it is difficult to process, and it takes a special effort to understand it. These differences between RBR/LBR and CER led numerous linguists to the belief that constructions containing higher-level CER are absent in human languages (e.g., Reich & Dell, 1977; Reich, 1969; Labov, 1973). While empirical research has shown by now that higher level CER constructions do occur in written and spoken language (for discussion see Sampson, 2001), it is generally accepted that these constructions are rarer than and judged as less grammatical than LBR/RBR at the same level of complexity. This raises an interesting problem for nativism. If recursion is at the heart of linguistic creativity and if language depends on an innate, genetically specified mechanism, then it is curious that different types of recursion pose different demands on language processing. Should we not expect that such closely related properties of language as LBR/RBR and CER are underwritten by very similar genetically specified mechanisms? This seemed indeed to be the default assumption of many nativists (e.g., Miller & Chomsky, 1963; Marcus, 1980; Church, 1982; Stabler, 1994; Chomsky, 2012). It was proposed that the problems with multiple CER arise not from linguistic but from psychological mechanisms (e.g., memory and attention span limitations, difficulties to paraphrase and fluently read sentences with multiple CER, for discussion see Sampson, 2001; Christiansen & MacDonald, 2009).

Recently the assumption that innate mechanisms underwrite the acquisition of multiple CER has been challenged (e.g., MacWhinney, 2004; Sampson, 2001; Christiansen & Mac Donald, 2009). Computational modeling of recursion has two distinct purposes. First, if it can be shown that non-domain specific models can imitate the performance regarding recursive abilities of human speakers, the assumption that a domain-specific innate faculty is required for recursion is challenged. Second, if it can be shown that models that are not limited by memory and other non-linguistic factors process and comprehend LBR/RBR and CER in a similar way to humans, then it appears plausible to suggest that the differences are of a linguistic nature.

Elman (1993) tested the hypothesis that “connectionist networks possess the requisite computational properties for modeling those aspects of natural language which are beyond the processing capacity of finite state automata [e.g. recursion]” (Elman, 1993, p. 75). From an artificial grammar he generated an input corpus of sentences with the following properties: (i) subject nouns and their verbs agreed for number; (b) verbs either required direct objects or optionally permitted direct objects or precluded direct objects; and (c) sentences could contain multiple embeddings in the form of relative clauses and subordinate clause (for full details see Elman, 1993, pp. 75 - 77). The network was trained to take one word at a time and predict what the next word would be. To make correct predictions the network needs to represent internally grammatical dependencies of the input. Elman found that networks that were trained on input of slowly increasing complexity achieved high performance and “generalized to a variety of novel sentences which systematically test the capacity to predict grammatically correct forms across a range of different structures” (Ibid., p. 77). On the other hand, networks that had access to the entire input corpus at once performed very poorly. Finally, networks that had been exposed to the complete input corpus from the beginning but were given a slowly increasing memory capacity (for details see pp. 78-79) had a prolonged initial learning phase but performed very well after that.

Elman claims that children will probably neither encounter the first nor the second condition. That is the input is never carefully matched to the language acquisition stage the child is at (condition 1), and the child is never able to make use of all the information that is contained in the complete input (condition 2). But it is plausible to assume that when children are learning, they are in a similar position to that of the networks in condition 3: they are faced with complex input, but their ability to access the input is limited. Over time this ability improves just as in the networks where "the learning mechanism itself was allowed to undergo 'maturational changes' (in this case, increasing its memory capacity) during learning" (Ibid., p. 79).

There are two important points here. First, it could be shown that networks are able to 'learn' even such complex grammatical structures as CER from the input. This, again, indicates that a domain-specific mechanism may not be required to achieve this result. At the very least it shows that the statistical information that is present in the input could be sufficient for the acquisition of CERs. Second, the fact that networks that are 'handicapped' in some way (by exposure to limited input or limited memory capacity) are more successful 'learners' than networks that have access to the complete input corpus and maximal memory capacity from the beginning had not been predicted by the experimenter. It was actually necessary to perform the experiments to obtain "a deeper understanding of the principles which constrain learning in networks" (Ibid., p. 85). As our understanding of these principles improves, our ability to develop better computational models and relevant test procedures for children improves as well.

Christiansen (1994) trained SRNs on a recursive artificial language and found that performance differed for higher levels of complexity of RBR and CER. Christiansen and Chater (1999) attempted to model the human processing performance for RBR, CER and cross-dependency recursion (CDR). For this purpose they trained "connectionist networks on small artificial languages, which exhibit the different types of recursive structure found in natural language" (Christiansen & Chater, 1999, p. 159). They found that the networks performed well on RBR and single-embedded CER and CDR but that performance quickly degraded for CER and CDR when additional levels of embedding were added. Christiansen and Chater suggest that what "constrains the performance of the SRN appears to be architectural limitations interacting with the statistics of the recursive structures" (Ibid., p. 172). They also observed that, even though the models were not trained on constructions of recursive depth four, "there was no abrupt breakdown in performance for any of the three languages at this point... This suggests that these models are able to generalize to at least one extra level of recursion beyond what they have been exposed to during training" (Ibid., p. 182).

The performance of SRNs closely resembles that of humans, who also can process several levels of RBR but have difficulties processing doubly or more highly embedded CER and CDR. Christiansen and Chater (1999) observe that the difficulty of processing center-embedded structures is not confined to a linguistic context. They cite Larkins & Burns (1977), who demonstrated that when subjects were asked to name center-embedded pairs of letters and digits, they experienced the same difficulty as when processing center-embedded sentences. Christiansen and Chater hypothesize that non-linguistic processing constraints could be to blame for the poor performance in the CER/CDR tasks. The findings that SRNs, which are not language specific, show the same performance limitations could support this hypothesis. These results also have consequences for the nativist/empiricist debate. Christiansen and Chater propose that

> These results suggest a reevaluation of Chomsky's (1957, 1959) arguments that the existence of recursive structures in language rules out finite state and associative models of language processing. These arguments have been taken to indicate that connectionist networks, which learn according to associative principles, cannot in principle account for human language processing. But we have shown that in principle this argument is not correct: Connectionist networks can learn to handle recursion with a comparable level of performance to the human language processor. (Christiansen & Chater, 1999, p.199)

Overall, recent work in modeling has provided some reason for questioning the cogency of 'in principle' arguments against connectionism.

It might be argued that, nevertheless, some aspects of the models introduced so far are problematic. For example, the training procedures do not reflect the language-acquisition process in children, and the input may contain more examples of RBR, CER and CDR structures than input to which children are exposed. Christiansen and Chater do not claim that their model replicates the acquisition process in children. Their goal was to show that the statistical information contained in the input is sufficient for *some* mechanism to acquire the ability to process recursive structures. This goal has been achieved.

Other work has also shown that the language input contains sufficient implicit information to allow for the acquisition of recursive structures. Christiansen and MacDonald (2009) show that a connectionist model is capable of simulating human performance of processing complex center-embeddings in German and cross- dependencies in Dutch. As previously discussed, these recursive constructions are more difficult to process than the simpler, right- and left-recursive structures. The authors suggest, “the ability to process recursive structure is acquired gradually, in an item-based fashion given experience with specific recursive constructions” (Christiansen & MacDonald, 2009, p.127). The SRN was trained on an artificial context-free grammar with a 38-word vocabulary (for details see Ibid., pp. 130-132) and then tested on novel sentences. The SRN was able to acquire complex grammatical regularities and “to make nonlocal generalizations based on the structural regularities in the training corpus” (Ibid., p. 132). Another important finding was that for both humans and SRNs doubly embedded CER were harder to process than doubly embedded CDR. This confirms the finding of Christiansen & Chater, 1999). The close fit to human data also extended to novel predictions where the models made grammaticality judgments similar to those of human subjects (Christiansen & MacDonald, 2009, p.149). Obviously there are some limitations to these models. The vocabulary is very small, and only a small subset of grammatical regularities is covered. Whether or not future models can be scaled up to the full complexity of human language remains to be seen.

Models that simulate the conditions under which human children learn language perform similarly to children. When children acquire language, the processing and production of recursive structure emerges gradually over time (Dickinson, 1987). This is contrary to Chomsky's assumption (1980) that the processing and production of recursive structure is virtually instantaneous. The work of Christiansen and MacDonald (2009) showed that, just like human children, SRNs do have a learning curve and their performance improves over time. Overall computational modeling of the acquisition of recursive structure has shown that there are several parallels between the performance of children and SRNs. This fact alone, of course, does not prove that the mechanisms exploited by SRNs are the same as those used by children. But the findings discussed here indicate that an innate domain-specific mechanism is not necessary to account for the acquisition of these structures. And the results obtained can direct future research and, it is hoped, provide insights into the mechanisms that allow human children to acquire language.

4. Conclusions

I have discussed computational models of language acquisition and evaluated whether or not the criticisms proposed by Chomsky (2000, 2009) and McGilvray (2009) apply. Based on the models I discussed, I suggest that the main points of criticism are not justified. Connectionists do not hold that neural nets (and by extension the models they use to simulate those nets) are initially "undifferentiated, approximating Locke's blank slate" (McGilvray, 2009, p. 110). They acknowledge that there is some structure initially built into their models and that this structure is relevant to the performance of the models. What is at issue for them is not whether or not neural nets are initially structured (they are) but whether or not this structure supports domain-specific learning or learning across different domains.

Empirical work has shown, it is not necessary to rely on an explicit notion of phoneme to succeed in the word segmentation tasks. The statistical information present in the input is sufficient for the success of segmentation models. Computational models show that it is possible to apply what has been learned from the training set to the test set without the need of explicit knowledge of the underlying structural rules and that the models can perform *as if* they had learned rules based on exposure to the statistical information contained in the input alone. In some cases modeling had results that did not conform to the predictions made prior to modeling. In these cases the obtained results influenced future research. First the experiments were repeated. When the repetition confirmed the earlier findings the theoretical assumption on which the models originally had been based were adjusted.

In spite of some impressive successes, computational models are still a considerable distance away from simulating all aspects of language acquisition. Recently work has begun to simulate more complex aspects of language acquisition from multiple-cue integration (e.g., Christiansen, et al., 2010; Rytting et al., 2010; Christiansen & MacDonald, 2009; Onnis et al., 2009; Monaghan & Christiansen, 2008), and language acquisition in different languages (e.g., Freudenthal et al., 2010; Jaroz, 2010; Blanchard et al., 2010; Christiansen & MacDonald, 2009), word-sense disambiguation (e.g., Waterfall et al., 2010) to the construction of a complete, empirical, generative model of the learning of syntactic patterns (Waterfall et al., 2010). Yet, we are still a considerable distance away from any model that simulates the complete process of language acquisition. Much interdisciplinary research remains to be done before we can hope to answer the question whether or not it will be eventually possible to combine many of the current 'small scale' models or if the complexity of the task exceeds the ability of data-driven models.

References

Blanchard, D., Heinz, J., & Golinkoff, R. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language, 37,* 2010, 487 - 511.

Brent, M. & Siskind, J. The role of exposure to isolated words in early vocabulary development. *Cognition, 81,* 2001, 31–44.

Chater, N., Reali, F. & Christiansen, M. Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences, 106,* 2009, 1015-1020.

Chomsky, N. *Syntactic Structures.* The Hague: Mouton & Co, 1957.

Chomsky, N. Review of Skinner's Verbal Behavior. *Language, 12,* 1959, 89-95.

Chomsky, N. *Aspects of the Theory of Language.* Cambridge, MA: MIT Press, 1965.

Chomsky, N. *Cartesian Linguistics. A chapter in the history of Rationalist Thought.* New York: Harper & Row, 1966.

Chomsky, N. Problems of knowledge. In: Allen, J. & Van Buren, P. (Eds.). *Chomsky: Selected Readings London:* Oxford University Press, 1971.

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

Chomsky, N. *Essays on Form and Interpretation*. Amsterdam: North Holland, 1977.

Chomsky, N. The Linguistic Approach. In: M. Piatelli-Palmerini. (Ed.), *Language and Learning*. Cambridge, MA: Harvard University Press, 1980. pp.107-130

Chomsky, N. On cognitive structures and their development. A reply to Piaget. In: Piattelli Palmerini (Ed.) *Language Learning: A debate between Noam Chomsky and Jean Piaget*. Cambridge MA: Harvard University Press, 1983. pp. 31-59

Chomsky, N. *Knowledge of Language*. New York: Praeger Publishing, 1986.

Chomsky, N. Linguistics and adjacent fields: a personal view. In: Kasher, A. (Ed.). *The Chomkyan Turn*. Cambridge, MA: Blackwell, 1991. pp. 3 - 25

Chomsky, N. *Language and Thought*. Wakefield, RI: Moyer Bell, 1993.

Chomsky, N. *The minimalist program*. Cambridge: MIT Press, 1995.

Chomsky, N. *The Architecture of Language*. Oxford: Oxford University Press, 2000.

Chomsky, N. Three factors in language design. *Linguistic Inquiry, 36,* 2005, 1-22.

Chomsky, N. Language and thought: Descartes and some reflections on venerable themes. In: Brook, A. (Ed.). *The prehistory of cognitive science*. New York: Palgrave Macmillan, 2007. pp. 38- 66

Chomsky, N. Cognition and Language. In: Oezsoy, A. & Nakipoglu, M. (Eds.) Noam Chomsky on Language and Cognition. Munich: Lincomb GmBH, 2009. pp. 5 - 21

Chomsky, N. The mysteries of nature how deeply hidden? In: Bricmont, J. & Franck, J. (Eds.) *Chomsky Notebook.* New York: Columbia University Press, 2010. pp. 3-33

Chomsky, N. *The Science of Language*. Cambridge: Cambridge University Press, 2012.

Christiansen, M., Dale, R., & Reali, F. Connectionist explorations of multiple-cue integration in syntax acquisition. In: S. Johnson (Ed.), *Neoconstructivism: The new science of cognitive development.*. New York: Oxford University Press, 2010. pp. 87-108

Christiansen, M. & MacDonald, M. A usage based approach to recursion in sentence processing. *Language Learning, 59 (Suppl. 1),* 2009, 126-161.

Christiansen, M., & Kirby, S. (Eds.). *Language Evolution*. Oxford: Oxford University Press, 2003.

Christiansen, M. & Chater, N. Connectionist psycholinguistics in perspective. In M. Christiansen & N. Chater (Eds.), Connectionist psycholinguistics. Westport, CT: Ablex, 2001. pp. 19-75

Christiansen, M. & Chater, N. Toward a Connectionist Model of Recursion in Human Linguistic Performance. *Cognitive Science, 23,* 1999, 157-205.

Christiansen, M. & Allen, J. Coping with variation in speech segmentation. In A. Sorace, C. Heycock & R. Shillcock (Eds). Proceedings of the GALA ‘ 97 conference on language acquisition : Knowledge representation and processing. Edinburgh: Edinburgh University Press, 1997. pp. 327–32

Christiansen, M. *Infinite languages, finite minds: Connectionism, learning and linguistic structure*. Doctoral dissertation, University of Edinburgh, 1994.

Crain, S. Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14*, 1991, 597–650.

Crain, S. & P. Pietroski. Nature, nurture and Universal Grammar. *Linguistics and Philosophy* 24, 2001, 139-186.

Crain, S., & Pietroski, P. Why Language Acquisition is a Snap. *Linguistic Review, 19,* 2002, 163-83.

Dickinson, S. Recursion in development: Support for a biological model of language. *Language and Speech, 30,* 1987, 239–249.

Edelman, S. & Waterfall, H. Behavioral and computational aspects of language and its acquisition, *Physics of Life Reviews 4,* 2007, 253-277.

Ellefson, M., & Christiansen, M. Subjacency Constraints Without Universal Grammar: Evidence from Artificial Language Learning and Connectionist Modeling: In L. R. Gleitman & A. K. Joshi (Eds.), *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000. pp. 645-50

Elman, J. Finding structure in time. *Cognitive Science, 14,* 1990, 179-211.

Elman, J. Distributed representations, simple recurrent networks, andgrammatical structure. Machine Learning, 7, 1991, 195-224.

Elman, J. Learning and development in neural networks: Theimportance of starting small. *Cognition, 48,* 1993, 71-99.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press, 1996.

Fiser, J., & Aslin, R. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 2001, 499-504.

Fitch, T., Hauser, M., & Chomsky, N. The evolution of the language faculty: Clarifications and implications. *Cognition 97,* 2005, 179–210.

Freudenthal, D., Pine, J., & Gobet, F. Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language, 37,* 2010, 643- 669.

Gordon, P. Learnability and feedback: A commentary on Bohannon and Stanowicz. *Developmental Psychology, 26,* 1990, 215–218.

Hare, M., & Elman, J. Learning and morphological change. *Cognition, 56,* 1995, 61-98.

Hauser M., Newport E., & Aslin R. Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton-top tamarins. *Cognition, 78,* 2001, 53-64.

Hauser, M., Chomsky, N., & Fitch, W. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science, 298,* 2002, 1569-1579.

Jaroz, G. Implicational markedness and frequency in constraint-based computational models of phonological learning *Journal of Child Language, 37,* 2010, 565- 606.

Karaminis, T. & Thomas, M. A Cross-linguistic Model of the Acquisition of Inflectional Morphology in English and Modern Greek. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2010. pp. 730-735

Labov, W. The place of linguistic research in American society. In: E. Hamp, (Ed.) *Themes in Linguistics: the 1970s*. The Hague: Mouton, 1973.

Larkin, W., & Burns, D. Sentence comprehension and memory for embedded structure. *Memory and Cognition, 5*(1), 1977, 17-22.

Legate, J. & Yang, C. Empirical Re-Assessment of Stimulus Poverty Arguments. *Linguistic Review 19*, 2002, 151-162.

MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

MacWhinney, B. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language 31*, 2004, 883–914.

MacWhinney, B. Computational models of child language learning: an introduction. *Journal of Child Language, 37,* 2010, 477-485.

Marcus, G. Language acquisition in the absence of explicit negative evidence: Can simple recurrent networks obviate the need for domain-specific learning devices*? Cognition, 73,* 1999, 293-296.

Marcus, G. & Brent, I. Are There Limits to Statistical Learning? *Science, 300*, 2003, 53-54.

Marcus, M. *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press, 1980.

Maye, J., Werker, J., & Gerken, L. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* 2002, B101–B111.

McDermott, D. *Mind and Mechanism*. Cambridge, MA: MIT Press, 2001.

McGilvray, J. On the innateness of language. In. Stainton, R. (Ed.) *Contemporary debates in cognitive science*. Malden, MA: Blackwell Publishing, 2006. pp. 97 - 112

McGilvray, J. Introduction to the third edition. In: Chomsky, N. *Cartesian Linguistics. A chapter in the history of Rationalist Thought.* Cambridge: Cambridge University Press, 2009.

Miller, G. & Chomsky, N.Finitary models of language users. In: R. Luce, R. Busch, E. Galanter (Eds. Reich & Dell, 1977; Reich, 1969; Labov, 1973) *Handbook of mathematical psychology*. New York: Wiley, 1963.

Mintz, T. Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*, 2002, 678-686.

Mintz, T. Frequent Frames: Simple Co-occurrence constructions and their links to linguistic structure. In: E. Clark & B. Kelly (Eds.) Constructions in Acquisition. Stanford: CSLI Publications, 2006.

Monaghan, P. & Christiansen, M. Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data*. Amsterdam: John Benjamins, 2008. pp. 139-163

Monaghan, P., & Christiansen, M. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*, 2010, 545-564.

Newman, R. The cocktail party effect in infants revisited: Listening toone's name in noise. *Developmental Psychology 41*, 2005, 352–62.

Newman, R., Bernstein Ratner, N., Jusczyk, A., Jusczyk, P. & Dow, K. Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology 42,* 2006, 643– 55.

Onnis, L., Christiansen, M. & Chater, N. Connectionist models of language processing. In: L.R. Squire (Ed.) *New encyclopedia of neuroscience. Vol. 3*. Oxford, U.K.: Elsevier, 2009. pp. 83-90

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. Implicit learning in real world context: The case of orthographic regularities. *Journal of Experimental Psychology: General, 130,* 2001, 401-426.

Perfors, A., Tenenbaum, J., & Wonnacott, E. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language, 37,* 2010, 607-642.

Perruchet, P., & Vinter, A. The Self-Organizing Consciousness. *Behavioral and Brain Sciences, 25,* 2002, 297- 330.

Pinker, S. *Learnability and Cognition*. Cambridge, MA: MIT Press, 1989.

Pinker, S. *The language instinct*. New York: HarperCollins, 1994.

Pinker, S. Words and rules. *Lingua, 106,* 1998, 219-242.

Pinker, S., & Prince, A. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 1988, 73–193.

Plunkett, K., & Juola, P. A connectionist model of English past tense and plural morphology. *Cognitive Science, 23,* 1999, 463-490.

Reali, F., & Christiansen, M. Uncovering the richness of the stimulus: Structural dependence and indirect statistical evidence. *Cognitive Science*, 29, 2005, 1007- 1028.

Redington, M., & Chater, N. Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences, 1,* 1997, 273-281.

Redington, M., Chater, N. & Finch, S. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22,* 1998, 425-469.

Reich, P.. The finiteness of natural language. *Language* 45, 1969, 831-43.

Reich, P. & G. Dell . Finiteness and embedding. In: R. Di Pietro & E. Blansitt (Eds.), *The Third LACUS Forum, 1976,* Hornbeam Press (Columbia, South Carolina), 1977.

Rohde, D., & Plaut, D. Less is less in language acquisition. In P. Quinlan (Ed.), *Connectionist modelling of cognitive development*. Hove, UK: Psychology Press, 2003. pp. 189-231

Rumelhart, D. & McClelland, J.. On learning the past tense of English verbs. In J. McClelland, D. Rumelhart, and PDP Research Group (Eds), *Parallel distributed processing: Volume 2: Psychological and Biological models.* Cambridge, MA: MIT Press, 1986. pp. 216-271

Rytting, C., Brew, C., & Fosler-Lussier, E. Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language, 37,* 2010, 513-543.

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

Sampson, G.. *Empirical Linguistics*. London: Continuum, 2001.

Smith, N.. *Chomsky. Ideas and Ideals*. Cambridge: Cambridge University Press, 1999.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. Unsupervised learning of natural languages. *Proceedings of the National Academy of Science, 102,* 2005, 11629-11634.

Stabler E. The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, & K. Rayner. (Eds.), *Perspectives on Sentence Processing*, 1994. (pp. 303-336). Lawrence Erlbaum.

Terrace, H.. Chunking and serially organized behavior in pigeons, monkeys and humans. In R. Cook (Ed.). *Avian visual cognition* [On-line], 2001.
Available: www.pigeon.psy.tufts.edu/avc/terrace/

Waterfall, H., Sandank, B., Onnis, L., & Edelman, S.. An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language, 37,* 2010, 671-703.

Wonnacott, E., Newport, E. & Tanenhaus, M.. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology 56,* 2008, 165–209.