Neural Syntax

Hartmut Fitz

Neural Syntax

ILLC Dissertation Series DS-2009-02

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: `illc@science.uva.nl`
homepage: `http://www.illc.uva.nl/`

Neural Syntax

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D. C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen
in de Agnietenkapel der Universiteit
op woensdag 1 juli 2009, te 10.00 uur

door

Hartmut Fitz

geboren te Urach, Duitsland

Promotiecommissie

Promotor: prof. dr. M. van Lambalgen
Co-promotor: dr. F. Chang

Overige leden:
prof. dr. A. E. Baker
dr. K. R. Blutner
dr. S. Frank
prof. dr. M. J. B. Stokhof
prof. dr. F. J. M. M. Veltman

Faculteit der Geesteswetenschappen

Printed and bound by PrintPartners Ipskamp, Enschede.

ISBN: 978-90-5776-190-4

Contents

Acknowledgments

When I started my PhD with a specialization in logic and computability, I had no inkling that I would end up writing a dissertation on computational models of language acquisition. The transition was steep at times, but always gratifying, and to engage in experimental work has humbled my philosophical mind considerably.

First and foremost I would like to thank my supervisor Michiel van Lambalgen for allowing me to follow this path. He advised me to seek out my own private *sweets shop* filled with jars of excitement, and I am very grateful he did not let me get away with anything less. From his admirable work and his guidance I learned how the confining barriers of scientific disciplines can be broken down in research. This has been a truly liberating experience, one for which I cannot thank him enough. I also want to thank Michiel for his exceptional generosity and patience, and his support and encouragement whenever I was wavering.

The bulk of this thesis would not have been possible without the expertise of Franklin Chang. Franklin put up with my modelling inexperience without hesitation, and patiently let me tinker with his sophisticated model until I got the hang of it. Throughout my project he unreservedly shared his knowledge, and provided invaluable advice and assistance on every inch of the model-related work (and way beyond). I could not possibly have asked for a more invigorating and supportive co-promotor.

I thank my committee members—Anne Baker, Reinhard Blutner, Stefan Frank, Martin Stokhof, and Frank Veltman—for taking the time to stomach this king-sized manuscript.

I have profited a great deal from discussing aspects of my research with Reinhard Blutner, Rens Bod, Harald Clahsen, Gary Dell, Stefan Frank, Evan Kidd, Theo Marinis, Martin Stokhof, Henk Zeevat and Jelle Zuidema. I would like to mention Morten Christiansen in particular, whose stimulating work inspired several parts of my thesis. I want to thank him for extensive discussions and advice, his support with grant applications, for giving me the opportunity to present my work in his seminar at Cornell, and his hospitality during my visits to Leipzig and Ithaca.

I experienced the ILLC as a very diverse and vibrant community and I am thankful for the opportunity to pursue my PhD here. I want to thank Frank Veltman especially, who kindly agreed to fund conference visits after my appointment with the ILLC had ended. Many thanks to Ingrid van Loon, Tanja Kassenaar, Karin Gigengack and Marco Vervoort for having been helpful in countless ways.

I also want to thank everyone from the Cognitive Science Center Amsterdam. To me the annual CSCA summer school was like the fifth season and a great source of inspiration on many occasions.

Cheers to all my colleagues on the second floor of the Philosophy Department for creating such a pleasant work environment. For one reason or another I would like to thank in particular Tine Wilde, Marian Counihan, Fabrice Nauze and Boudewijn de Bruin. Furthermore, thanks to my friends abroad—Oliver Kutz, Sibylle Tröster, Michael Wagner and Frederik Witt—for being out there.

Although they have no idea they were involved in this thesis, thanks to Jon, Spencer, Don, Tom, Irene and Bethany for painting many of my late working hours in peculiar sonic colors.

A special thanks to Alexandra Silva whose fabulous dinner parties changed my life from worse to better during the final stages of writing.

And finally, my warmest thanks go to my mom and Josha for their tremendous support, and to Helle for being such a wonderful partner and friend.

Amsterdam Hartmut Fitz
April, 2009.

Chapter 1

Introduction

A key property of the human language faculty is the capacity to produce and comprehend a large and indefinite number of different expressions which are assembled from a fairly small inventory of memorized words. Natural language allows the combination of words into phrases and clauses and these can be organized into hierarchically structured complex sentences. In order to cope with this expressivity, language learners must be capable of generalizing acquired linguistic knowledge beyond their immediate experience. Explaining the nature of such generalizations and how they can be accomplished is a central endeavor for any theory of language acquisition.

1.1 Computational models in language acquisition

Theories of language acquisition are intimately tied to theories of language. Theories of language characterize the *acquirendum*, the kind of linguistic property or knowledge that is being learned and generalized. The processes of learning and generalization as such are then investigated by theories of acquisition. The reliance of theories of acquisition on theories of language is inevitable since a theory of acquisition needs to be informed by a theory of language about which aspects of language are general patterns and not merely rote forms. Thus, theories of language provide the conceptual and theoretical scaffold in the very description of the learning and generalization tasks studied by theories of acquisition.[1] As a consequence, the study of language acquisition inherits many assumptions about the *acquirendum* and the nature of a learner's linguistic knowledge from theories of language. In English, for example, subject-auxiliary inversion occurs in a variety of utterance types such as yes/no-questions (*Did they win?*), counterfactual conditionals (*Had he left five minutes later, he would have missed the train*), exclamatives (*Wow, does that taste good!*), comparatives (*He has ventured further than have his contemporaries*), and several others. Subject-auxiliary inversion

[1] As Pinker (1990) has put it succinctly, "to understand how X is learned, you first have to understand what X is."

is regarded as a purely syntactic generalization in approaches to language which emphasize the autonomy of syntax (Newmeyer, 2000). In constructivist approaches, on the other hand, subject-auxiliary inversion is characterized as a grammatical category held together by semantic and pragmatic functional similarity (Goldberg, 2006). In the former framework, what must be explained by a theory of acquisition is how (and whether) a single, all-encompassing grammatical movement rule can be induced from linguistic input. According to the latter view, it must be explained how (and whether) a grammatical category can be abstracted from lexically-specific patterns of inversion by means of 'functionally-based distributional analysis' of the input (Tomasello, 2003). Thus, both characterizations of the generalization task are shaped by assumptions from distinct theories of language, and these assumptions enter into the formulation of experimental hypotheses and design, and often determine in a subtle way how behavioral data should be interpreted.

In this way, language acquisition theory is strongly influenced by linguistic theory; a theory of language isolates and describes properties of natural language that a learner must come to know, a theory of acquisition collects behavioral data and evaluates the stipulations of linguistic theory against this data.[2] In the above example of subject-auxiliary inversion, a movement-rule account might predict category-general knowledge and the absence of errors in children's data. Children, however, make inversion mistakes and movement accounts need to be repaired by explaining errors in terms of several rules and their occasional misapplication. Moreover, Rowland and Pine (2000) found that children make lexically-specific errors in subject-auxiliary inversion which, arguably, contradicts the idea of knowing a category-general movement rule. Their analysis suggests that children learn subject-auxiliary inversion in *wh*-questions item by item and develop more abstract generalizations later on. The functional-constructivist approach to subject-auxiliary inversion, to which Rowland and Pine subscribe, is consistent with their data. Conversely, the data is inconsistent with an *acquirendum* posited by a specific theory of language—i.e., subject-auxiliary is governed by movement rules—which might thus be rejected in this domain. An explanatory relationship is established between a linguistic theory and acquisition data by this kind of analysis. But this relationship is conceptual in nature, not causal. It does not answer the crucial question how adult-like generalizations can be acquired based on linguistic experience. It also does not answer whether the functionally defined category of inversion constructions is psychologically real in language processing. Most importantly, a constructivist characterization of the *acquirendum* by itself does not enable quantitative predictions which could be tested independently.

These shortcomings result from the failure to specify the *mechanisms and processes* subserving the acquisition and generalization of inversion. In other words, there can be no satisfactory explanation growing out of a theory of acquisition plus developmental data, without an account of how the human language processor is *affected* by the

[2] Of course this picture is a dramatic oversimplification of the often intricate reciprocal relationship of linguistic theory and theories of acquisition in natural language research.

properties of the input language according to linguistic theory (e.g., the functional similarity of inversion constructions), and how this processor *causes* the observed data. This is, I believe, the main reason why we need to supplement theories of acquisition with the study of learning and generalization in the framework of computational models (Figure 1.1). Computational mechanisms of learning and generalization can establish an explanatory link between a theory of language and acquisition data by providing answers to *how* questions. How do properties of natural language (according to linguistic theory) affect the human processor and how does the processor cause the observed behavior in acquisition? Rowland and Pine argue, for instance, that the error patterns they observed in children are best explained by the frequency of inversion patterns in the learning environment which give rise to lexically-specific knowledge. This claim is difficult to test in a verbal theory of acquisition, but it could be tested in a computational learning model which is sensitive to distributional properties of the input. Such a model could help bridge the gap between input data and observed developmental data and thus validate a specific theory of acquisition. Moreover, computational models are formally precise, consistent theories themselves, which do not leave components unexplained or underspecified (else the model could not produce any useful behavior). These properties are particularly desirable when due to their inherent vagueness several verbal theories appear to be in line with some acquisition data. On the downside, computational models are often highly simplifying and can not implement verbal theories precisely and in every detail. Hence, they usually do not cover the full range of data explained by theories of acquisition. Unlike such theories, however, computational models allow novel, quantitative predictions which in turn can be tested in behavioral experiments and this can help justify a particular theory of acquisition.[3]

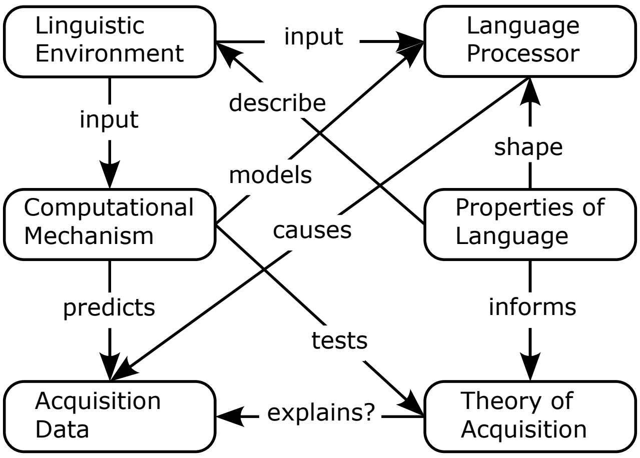


Figure 1.1: The role of computational models in the study of language acquisition.

[3] Models are not only heuristic tools in the justification of theories but also in their very discovery (Gigerenzer, 2000).

1.2 Why neural networks?

When motivating the use of neural networks over other kinds of architectures, frequently the particular strengths of connectionist systems in modelling aspects of human cognition are invoked. For instance, neural networks deal well with noisy input and local malfunction, causing ‘graceful degradation’ instead of catastrophic failure. They can learn graded category membership, attend to subtle statistical regularities, satisfy multiple, conflicting constraints, and so forth. Although these properties clearly are advantageous in modelling language processing, there are other, more mundane reasons to study syntactic development with neural networks: these models can learn from natural language input and they develop syntactic representations in an autonomous, self-organizing manner. As was argued above, in verbal theories of language acquisition, it is difficult to describe generalization tasks of a human learner without adopting a specific theory of syntax to characterize the task itself. Statistical learning models such as neural networks allow us to study the mechanisms of linguistic generalization in a less theory-dependent way. This is because they learn from language corpora by domain-general algorithms. No explicit, language-specific programming is required, the ‘program’ of the model is found by adjusting its free parameters adaptively. Specifying the model’s learning environment, input/output encoding, and learning procedures does not involve theory-laden assumptions about the syntax underlying the target language. Neural network models ‘find’ syntactic representations autonomously in the process of generating a solution to a computational problem and these representations are not preconceived by the experimenter nor do they necessarily map onto the syntactic categories postulated by descriptive linguistics. Because of domain-general learning and the autonomy of representations, neural network models *prima facie* are ideally suited for modelling syntactic development and sentence processing.

Neural networks are often advertised for their neurobiological plausibility. These models are an attempt at emulating information processing in the human central nervous system. It should be pointed out, however, that the artificial systems studied in this thesis are perhaps no more neurobiologically accurate models of the brain than the Dutch telephone system, or the World Wide Web with its ‘highly interconnected’ computing units and ‘massively parallel’ flow of information. This is because artificial neurons and their connection weights do not adequately model the biochemical and bioelectric properties of living neurons and their synaptic connectivity (see also Section 2.3.3); nor do they reflect the variety of cell types in the nervous system. But computational models should not seek to replicate reality in all its vast complexity. Rather they should make helpful abstractions and reasonable simplifications to isolate critical aspects of reality. One such aspect is the way in which information is processed concurrently by simple computational units through activation spread and signal transformation, without central control, and without manipulating explicit data structures. Control is achieved through communication and coordination between individual cells. This biological ‘model of computation’, although implemented in networks of highly simplified neurons, I believe, captures the essence of computational processes in the human brain.

Yet, whether abstracting away from a myriad of other, more specific properties of real neurons and how they exchange information is appropriate for artificial networks to yield cognitively plausible linguistic behavior is an entirely different question. It is one that can only be addressed experimentally, and the present thesis hopefully contributes to investigating this issue.

While neural networks are widely considered suitable models for 'low-level' cognitive functions (such as memory, attention, perception, recognition, category learning, and motor control), they are also widely considered ill-suited for modelling 'representationally-intense' cognitive functions such as language processing, reasoning and decision-making. In the literature, neural networks have been heavily criticized specifically for their inability to generalize linguistic knowledge in human ways—due to, e.g., the nature of their syntactic representations (Fodor and Pylyshyn, 1988), their overreliance on statistical information (Marcus, 1998, Marcus et al., 1999), and their failure to behave systematically (Hadley, 1994, 2004, Fodor and McLaughlin, 1990). Furthermore, it has been argued that neural networks generalize in empirically incorrect ways (Pinker and Prince, 1988) and that they are downright unsuitable for explaining syntactic generalization, because they do not represent syntax at all.[4] The present thesis partially grew out of my dissatisfaction with these negative verdicts which are often based on specific neural network models and might not extend to all connectionist systems. I aim to show that the Dual-path model, studied in this work, displays interesting generalization behavior which makes it a suitable computational platform for investigating human syntactic development and sentence processing.

1.3 Thesis outline

This manuscript is organized as follows. In Chapter 2, I begin giving a brief overview of results that characterize the properties of neural networks as mathematical objects. By identifying their computational capacities these models can be related to the formal complexity of artificial string languages. The issue of learning such languages from finite data will be considered before I turn to studies which replaced string languages with more naturalistic input. In both domains I focus mainly on work with the simple-recurrent network model of Elman (1990, 1991) which has been used extensively in studying the acquisition of complex sentence structure, although other models will be discussed too. In particular, I will review the seminal approach of Christiansen (1994) and Christiansen and Chater (1999b), to provide a neural network account of recursion in human performance. The chapter serves to map out the conceptual landscape in which the current work is located.

Although the simple-recurrent network model yielded important insights into how complex syntactic structure could be learned from temporally extended data, I will argue that the model is limited in a number of ways. The construction and use of meaning

[4] Harald Clahsen in personal correspondence, June 2007.

is essential in child language acquisition and adult processing and neural network models need to incorporate this dimension of human linguistic behavior. Enriching such models with semantic information transforms the computational learning problem from grammar induction into meaning-form transduction. The Dual-path model is a model of sentence production and syntactic development which is able to represent sentence meaning and incrementally map it onto a sentence form (Chang, 2002; Chang, Dell, and Bock, 2006). It learns from exposure to sentences paired with their meaning. I explain the architecture of the Dual-path model in Chapter 3, motivate critical assumptions behind its design, and discuss past research using this model.

Chapter 4 describes and compares several extensions of the basic Dual-path model to accommodate the processing of multi-clause utterances. These extensions are evaluated against computational *desiderata*, such as good learning and generalization performance and the parsimony of input representations. A single-best solution to encoding the meaning of complex sentences with restrictive relative clauses will be isolated, which forms the basis for all subsequent simulations. This chapter should be viewed as establishing necessary architectural preliminaries rather than containing novel insights into human sentence processing.

After determining suitable semantic representations for complex events which allow the Dual-path model to produce sentences with embedded clauses, I analyze the model's learning dynamics in detail in Chapter 5. In the first part of this chapter, I examine the model's differential performance on distinct structures in the input. Syntactic alternations prove to be particularly difficult to learn because they complicate the meaning-to-form mapping the model has to acquire. In the second part, I probe the internal representations the model has developed in learning the target language. By inspecting activation patterns at individual layers during processing, it can be demonstrated that the model induces grammatical categories over word sequences, and assigns thematic roles to sentence constituents incrementally. I also investigate whether the model represents phrase structure, verb argument structure, and the clausal structure of complex utterances. It is argued that traditional phrasal categories are not detectable but that the model acquired the argument structure of the construction types in its input language. Furthermore, it is shown that the Dual-path model represents the hierarchical organization of distinct multi-clause utterances.

The juice of this dissertation is contained in Chapters 6–8. In Chapter 6, the Dual-path model's generalization capacities are put to the test in a variety of tasks. First, I show that the model's syntactic representations allow the transfer of knowledge between clauses, which is a precondition for generalizing basic constructions to more complex constructions. The artificial language and corresponding meaning representations are extended to sentences with up to four nested relative clauses. I demonstrate that the Dual-path model can assemble novel multi-clause utterances with several embeddings from experience of simpler structures. In other words, the model generalizes structurally. Semantic similarities in the conceptual structure between novel and familiar sentence types play a critical role in this task. It is an important human capacity to learn words in one syntactic/semantic context and use them in another. The Dual-path model is able to generalize familiar lexical items to novel thematic roles, in novel constructions, at novel levels of embedding. Since most of these constructions are not experienced in learning, this property might be called *super*-strong semantic systematicity (Hadley, 1994). Humans can also comprehend and produce utterances with a novel hierarchical organization, they are recursively productive. According to Hauser et al. (2002), recursion is a core capacity of the human language faculty. I identify learning conditions under which the Dual-path model displays recursive productivity. It is shown that the model's behavior is consistent with human behavior in that production accuracy degrades with depth of embedding, and right-branching recursion is easier for the model than self-embedding recursion.

A central issue in language acquisition concerns the question which syntactic constructions can be learned from experience and which require some kind of biological endowment. Complex polar interrogatives—yes/no-questions with relative clauses—occupy a prominent place in this *nature versus nurture* controversy. They appear to be largely absent from child-directed speech and therefore necessitate innate structure-dependent principles for their learnability. In Chapter 7, I address this issue in the framework of the Dual-path model. I show that the model can induce the syntax of complex polar interrogatives piecemeal from simpler and similar structures which are warranted in a child's linguistic environment. In the absence of positive exemplars of the target structure in the input, both lexical and structural generalization are critical features of the Dual-path model in this task. The model's errors are compared with child language data and I argue that the model does not entertain erroneous syntactic hypotheses which would require overt correction or an innate learning bias. As a consequence, the orthodox formulation of the learning problem that children supposedly face might be ill-conceived. Since the model does not implement a traditional kind of language-specific universal grammar, these results are relevant to the *poverty of the stimulus* debate.

English relative clause constructions give rise to similar orderings of differential processing in adult comprehension (Keenan and Hawkins, 1987) and language production in development (Diessel and Tomasello, 2005). This pattern matches the typological universal called the *noun phrase accessibility hierarchy.* I propose an input-based explanation of this data in Chapter 8. The Dual-path model displayed the ordering of the hierarchy in syntactic development when exposed to plausible input distributions. But it was possible to manipulate and completely remove this ordering by varying properties of the input from which the model learned. This indicates, I argue, that patterns of interference and facilitation among input structures can explain the accessibility hierarchy in processing and development when all structures are simultaneously learned and represented over a single set of connection weights in a neural network model.

In Chapter 9, finally, I draw conclusions from this work, address some unanswered questions, and give a brief outlook how this research might be continued. An abstract precedes each chapter to help the reader keep track of the agenda.

Chapter 2

Recursion in neural networks

In this chapter I review previous research investigating the capacity of neural networks to represent and learn recursive structure. First, an overview of formal results is given regarding the mathematical properties of widely used neural network architectures. Then I review the application of such models in the domain of grammar induction for string languages of various complexity. Several computational studies of language learning and processing are discussed which aimed at showing that neural network models can cope with aspects of complex sentence structure found in natural language.

2.1 Introduction

When attempting to model natural language processing in neural networks—in particular the acquisition of recursive syntax—it is a question of immediate interest whether these models are computationally capable of coping with the amount of structural complexity attributed to natural language in a formal sense. Historically, negative results on the computational capabilities of neural networks have led to a decline of interest in these models, for instance, when Minsky and Papert (1969) showed that perceptrons (the simplest kind of feed-forward network with two layers and a threshold activation function) could not compute functions which are not linearly separable (e.g., the XOR-function). Whether other neural network types are computationally adequate for natural language processing is a complicated issue, the intricacies of which are often underestimated in the 'symbolism-versus-connectionism' debate. The spectrum of claims made in this debate ranges from downright rejection because "connectionist networks just don't compute the right kind of functions" to accommodate for natural language syntax[1] to declarations of triviality that "it is a simple matter to prove that neural networks can do anything that symbolic processors can do".[2] These quotes indicate that there is a considerable amount of uncertainty regarding the computational power of

[1] Ted Briscoe in discussion at the First Scottish-Dutch Workshop on Language Evolution, University of Amsterdam 2005.

[2] James Garson in the Stanford Encyclopedia of Philosophy entry on "Connectionism".

neural networks. This might be due, partially, to the vast diversity of neural network architectures that have been proposed, which impedes a straightforward computational analysis.

2.2 The computational power of neural networks

In the following sections I will give a very brief overview of formal results which have been obtained in recent years for multi-layer feed-forward and recurrent networks which, arguably, are the most important architectures used in connectionist natural language processing.

2.2.1 Feed-forward networks

Minsky & Papert speculated that their limitative result for perceptrons would also hold for multi-layer feed-forward networks which was proven wrong in that networks with additional layers are strictly more powerful (Grossberg, 1973; Rumelhart et al., 1986). Since then it has been shown that these models can be viewed as 'universal function approximators' in the following sense: let f be any continuous real function from $K \subset [0, 1]^n$ to $(0, 1)$. Then, for any $\epsilon > 0$ there exists a multi-layer feed-forward network with sigmoid activation function such that

$$E = \int_K |f(x) - \tilde{f}(x)| dx < \epsilon \tag{2.1}$$

where $\tilde{f}$ is the function computed by the network and E is the total approximation error. Thus, f can be approximated by such a network with arbitrary precision. A particularly simple, constructive proof of this proposition can be found in Rojas (1996), and this result is based on work by Hornik et al. (1989) and Funahashi (1989). Castro et al. (2000) have recently extended this result to continuous real functions $f : K \subset \mathbb{R}^n \to (0, 1)^m$ and networks with arbitrary continuous, strictly increasing squashing activation functions. This is a strong statement about the computational properties of these networks, but it should not be misconstrued as asserting or entailing that any such function f can be computed by a feed-forward network. For any degree of precision ϵ, a network can be constructed which approximates f to this degree. But for a better approximation $\epsilon' < \epsilon$, more hidden units must be invested; no single network is sufficient to approximate f to any desired degree of accuracy. Thus, proposition (2.1) assumes that networks can be constructed using unbounded resources. Moreover, the procedure for determining the parameters of a network (hidden units, weight size, etc.), which may be called the 'learning problem', is not efficient but NP-complete (see Rojas, 1996). Nonetheless, (2.1) indicates that the functional relationships which multi-layer feed-forward networks can represent are very complex. Consequently, it should be considered an *analytic principle of connectionist language processing* that any failure of such models in application must be due to inappropriate learning, an inadequate 'phenotype' (e.g., too few hidden units), or the lack of functional relations between input and targets in the training set (cf. Hornik et al., 1989).

2.2.2 Recurrent networks

Natural language exhibits many relationships between sentence constituents, for instance noun-verb agreement, pronominal and anaphoric binding, co-reference, verb-argument structure, and quantifier scope. These relationships are often complicated through complex syntactic constructions such as relative clauses, complement clauses, and conjunctive subordination/coordination, creating long-distance dependencies. Because of its hierarchical clause structure, language is more than a linear arrangement of words. But sentences are processed sequentially over time. It is therefore a major challenge for any model of natural language processing to explain how complex relationships between constituents can be represented and learned from temporally extended data. One way to approach this problem is by representing time 'spatially' through storing the sequential input in a separate memory system. Recurrent neural networks, on the other hand, track dependencies in temporal data through feedback connections. This recurrence creates a dynamic short-term memory of the previous activation states of units in the network. In that these states influence the network's input-output mapping, the temporal properties of sequentially presented data become encoded in the network's internal representations during learning. Thus, time is represented implicitly "by the effect it has on processing and not as a separate dimension of the input" (Elman, 1990, p. 180). As opposed to other more limited architectures, recurrent networks therefore lend themselves as particularly suitable for modelling aspects of natural language processing.

The recurrent neural networks (RNN for short) I consider here are n-dimensional dynamical systems over a bounded n-cube of reals. Each of the units x_j $(1 \leq j \leq n)$ assumes an analog-state in $[0, 1] \subset \mathbb{R}$ at discrete points in time $t \in \{0, 1, 2, \ldots\}$. The interconnections between units are given by the set of weights $\mathcal{W} = \{w_{ij} \in \mathbb{R} \,|\, 1 \leq i, j \leq n\}$ where w_{ij} means that unit x_j projects to unit x_i via a synaptic connection. Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ be bias constants, $\chi : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ an excitation function, and $\sigma : \mathbb{R} \rightarrow [0, 1]$ an activation function. Initially, at time $t = 0$, the network is placed in state $\mathbf{x(0)}$ which possibly includes external input at some units. The network is updated synchronously in parallel. The global network state $\mathbf{x(t)} = (\mathbf{x_1(t)}, \ldots, \mathbf{x_n(t)}) \in [0, 1]^n$ for all discrete time instants $t = 0, 1, 2, \ldots$ is computed as follows: each unit x_j collects the input $x_i(t)$ from all units such that $w_{ji} \in \mathcal{W}$ and computes its subsequent state according to the equation

$$x_j(t + 1) = \sigma(\chi_j(x_i(t), w_{ji}) - \theta_j). \tag{2.2}$$

If the excitation function χ_j is a linear combination of inputs for all units x_j, such as the common dot product $\chi_j = \sum_{i=1}^{n} w_{ji} x_i(t)$, the network is called *first-order*. θ_j in (2.2) is a weighted local input from a special constant bias unit. Next, the activation function σ is applied to all units' excitation level, determining the network's next global state $\mathbf{x}(\mathbf{t}+\mathbf{1})$. Usually, σ is a threshold function in Boolean networks, and a sigmoid function, such as $\tanh$, the logistic or the saturated-linear function in analog-state networks. If the same activation function σ is used for all units, the network is called *homogenous*. Recurrence is realized in this model in that weights in $\mathcal{W}$ are allowed to connect units with themselves, recurrent layers are realized by weights between units in the same layer. Generally, time-delayed connections can occur in RNNs at any of the network's units. In simple-recurrent networks (SRN) recurrence is limited to one hidden layer (Elman, 1990, 1991). The activation spread of a recurrent network can be visualized by unfolding the temporal delay into space. Thus, for fixed-length input sequences, a dynamical RNN can be mimicked by a larger, static feed-forward network (see Figure 2.1). This property is utilized in the backpropagation-through-time learning algorithm (BPTT, Rumelhart et al., 1986), which is an approximation to ideal gradient-descent error correction, where blame for a current mismatch would be assigned to the weights by taking into account the entire input history of the network. But unlike feed-forward networks, recurrent networks are capable of performing general computations for inputs of varying length.

Elman-type SRNs typically employ the real-valued, continuous logistic activation function

$$\sigma(\chi_j) = \frac{1}{1+e^{-\chi_j}} \tag{2.3}$$

where χ_j is the excitation of unit x_j. However, neural networks are implemented on systems which use fixed precision arithmetic, e.g., digital computers, so that the network's activation state and weight memory is finitized. Kremer (1995) has shown that under this assumption SRNs are computationally equivalent to deterministic finite-state machines (FSM).[3] For every such FSM there is an SRN which emulates it. Hence, in a representational sense SRNs are as powerful as any digital computer with finite memory. Whether for a given FSM this SRN can be found efficiently through learning is a different issue, which depends on the training conditions, learning algorithm and network topology.

Heterogenous, second-order RNNs with unbounded precision (unit output and weights) were shown to be Turing-equivalent by Pollack (1987). Siegelmann and Sontag (1991) considerably strengthened this result by demonstrating that there is a finite, homogenous first-order RNN $\mathcal{N}_{TU}$ with saturated-linear activation function and rational weights which is Turing-universal.[4] If σ is a threshold function, such RNNs coincide with FSMs. Thus, contrary to intuition, the multiplicative excitation function used in Pollack (1987) turned out to be unnecessary.[5] The proof in Siegelmann and Sontag (1991)

[3] The proof in Kremer (1995) considers only SRN with threshold activation functions but has been extended to sigmoids by Alquézar and Sanfeliu (1995).

[4] This activation function is piecewise linear $\sigma(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$.

[5] Modulo polynomial speed-up, multiplicative networks are computationally equivalent to first-order

is constructive and Neto et al. (1997) provided a construction method for modular RNNs to compute arbitrary partial recursive functions $\psi : \mathbb{N} \to \mathbb{N}$. Siegelmann (1999) constructed an RNN $\mathcal{N}'_{TU}$ which real-time simulates a universal Turing machine, i.e., it respects time complexity as well. These results show that under standard idealizations made in symbolic computation, such as unbounded time and memory, RNNs are equivalent to classical computational paradigms.

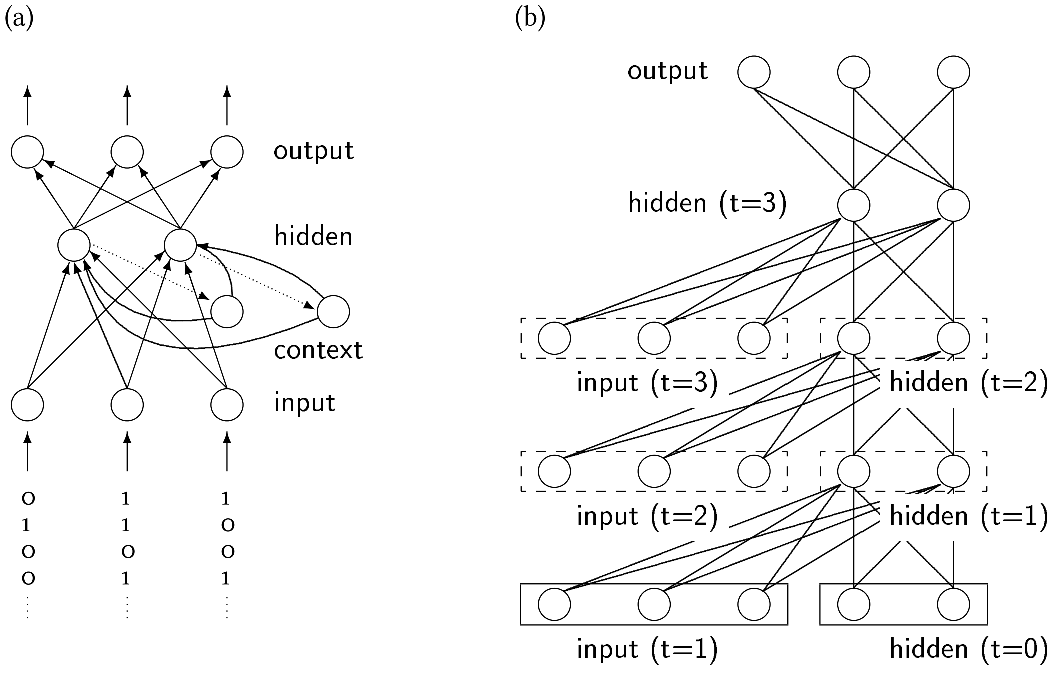


Figure 2.1: (a) An SRN with two hidden units, dotted lines indicate non-learnable copy-back connections which implement recurrence. (b) The same network unrolled for three time-steps.

A sigmoid activation function σ injects non-linearity and noise into the network because no matter how extreme a unit's excitation level, its response will never be binary. Interestingly, the RNNs described above remain Turing-universal under such conditions, confer the proof in Kilian and Siegelmann (1996). Noise becomes highly detrimental, however, if it affects a network in the sense of slightly 'shaking up' its global state during activation spread. Maass and Orponen (1998) showed that RNNs collapse to the computational capacities of FSMs when subjected to "any reasonable type of analog noise, even if their computation time is unlimited and if they employ arbitrary real-valued parameters" (p. 1082). In noiseless environments, such as pure mathematics, one may wonder what capacities real weights add over rational, i.e., finitely describable RNNs. Granted exponential time, real-weight RNNs are omniscient in that they recognize arbitrary languages over a finite alphabet. Restricted to polynomial time and with a

RNNs (confer Chapter 10 of Siegelmann, 1999). Consequently, higher-order networks do not give rise to a computational hierarchy.

linear-precision bound on all internal coefficients, they recognize all and only languages in the non-uniform complexity class P/poly.[6] Hence these RNNs remain *super-Turing* even when internal arithmetic operations are dynamically truncated as a function of the input length (see Chapter 4 in Siegelmann, 1999).

To summarize this brief and highly selective survey, the representational capacities of recurrent neural networks are very strong in terms of automata-theoretic notions of computability. Under real-world restrictions such as finite precision and finite memory, i.e., in any experimental application, RNNs are equivalent to the class of deterministic finite-state automata (which are an expressive subclass of finite-state machines). Thus, such RNNs can implement any realizable finite memory digital computer. This is also true in ecologically realistic conditions such as the presence of cortical noise. Under idealized conditions, i.e., as mathematical objects, RNNs are Turing equivalent or even super-Turing, depending on more fine-grained assumptions about the nature of parameters in these networks as sketched above.

2.3 RNNs and string language learning

To determine the computational properties of RNNs in a mathematically precise fashion is important to preempt objections against connectionist language processing based on architectural limitations. A main appeal of neural network models, however, lies in the capacity of these systems to *learn* from examples. In learning, the network configuration is altered by gradually adjusting the weights in such a way as to enable the network to map inputs to desired outputs in the training set. At the same time, the network may acquire dispositions to respond to novel input retaining the transformations encountered during training. The representations developed by the network through learning may enable it to *functionally* maintain structured information from the training environment and generalize it to novel stimuli. It is an issue of considerable practical interest whether recurrent neural networks can deal with the structural complexity of natural language in this sense of functional learning. While RNNs may in principle be computationally suitable to perform mappings required for natural language processing, learning this task from examples may be difficult or even impossible.

The degree of success in approaching this issue will depend on learning parameters, RNN topology, and critically on the learning task itself, i.e., the structural complexity of the target language. A measure of complexity is provided by the Chomsky hierarchy which is a containment ordering of classes of phrase structure grammars that generate formal languages of distinct automata-theoretic complexity (Figure 2.2). The task of learning string languages from the Chomsky hierarchy with neural networks can be explicated in several ways, the most general definition has been given in terms of the notion of a *dynamical recognizer* (Pollack, 1991; Moore, 1999). A dynamical recognizer $\mathcal{D}$

[6] P/poly is the class of languages recognized by a polynomial-time Turing machine with a polynomially bounded advice function and contains every undecidable unary language. It can be considered a class of 'efficient non-computability'.

Chomsky Hierarchy			
Grammar	Language	Automata	Rewrite Rules
Type-0	Recursively enumerable	Turing machines	$\alpha \to \beta, \alpha \neq \epsilon$
Type-1	Context-sensitive	Linearly-bounded, non-deterministic Turing machines	$\alpha A \beta \to \alpha \gamma \beta, \gamma \neq \epsilon$
Type-2	Context-free	Non-deterministic pushdown automata	$A \to \gamma$
Type-3	Regular	Deterministic finite-state automata	$A \to \alpha$ and either $A \to B\alpha$ or $A \to \alpha B$, α terminal

Figure 2.2: The Chomsky hierarchy of formal languages. Uppercase letters represent non-terminals, greek letters represent strings of terminals and/or non-terminals, ϵ is the empty string.

for a language $\mathcal{L}$ over a finite alphabet $\mathcal{A}$ consists of a space $\mathbb{R}^k$, a set $\mathcal{F} = \{f_a : \mathbb{R}^k \to \mathbb{R}^k \,|\, a \in \mathcal{A}\}$ of maps, an initial state $x_0 \in \mathbb{R}^k$ and a subset $\mathcal{H} \subset \mathbb{R}^k$ called accepting set. Let w be a word from $\mathcal{A}^*$, w_i its i-th letter and $|w|$ the length of w. Define the compositional map $f_w := f_{w_{|w|}} \circ \ldots \circ f_{w_2} \circ f_{w_1}$ and let $x_w := f_w(x_0)$. The language recognized by $\mathcal{D}$ is defined as $\mathcal{L}_\mathcal{D} = \{w \,|\, x_w \in \mathcal{H}\}$. In other words, $\mathcal{D}$ recognizes a set of words from $\mathcal{A}^*$ if the iteration of maps f_{w_i} for successive letters w_i of each w in the language $\mathcal{L}$, starting from an initial point x_0, terminates in a designated region $\mathcal{H}$ of $\mathbb{R}^k$.[7] Pollack (1991) demonstrated that a higher-order, cascaded recurrent network, when trained with backpropagation, was able to induce a dynamical recognizer for a depth-limited balanced parentheses language. This relation between string acceptance and neural networks as dynamical recognizers suggests a definition of discrete-time RNNs as neural-state Mealy or Moore machines (Carrasco et al., 2000; Carrasco and Forcada, 2001).[8] After an input string has been processed in its entirety, the RNN signals acceptance (or rejection) in a designated region of the output space (classification task). Alternatively, RNN can be trained to predict the next word in a string, including an end-of-word marker (prediction task).[9] For instance, the Elman-type SRN can be defined as a neural-state Moore machine and trained to accept all and only the strings w from some regular language $\mathcal{L}$ (see below).

[7] Note that the set-theoretic 'complexity' of $\mathcal{H}$ co-determines the capacity of $\mathcal{D}$, cf. Moore (1999).

[8] Moore and Mealy machines are equivalent types of finite-state automata with output, the former computes outputs over states, whereas the latter computes outputs over state transitions.

[9] The prediction task is psycholinguistically more plausible than the classification task which can be viewed as a grammaticality judgement.

2.3.1 Regular languages

A large number of studies have investigated the learnability of regular languages by various types of recurrent networks from finite data (commonly referred to as 'grammar induction'). The SRN model, for instance, was tested on this task by Cleeremans et al. (1989) and Maskara and Noetzel (1992) used a more potent variant of the SRN in which the network also had to produce the previous context and current input at each point in time. More general first-order RNN architectures have been studied by Manolios and Fanelli (1994) and Sanfeliu and Alquézar (1994), second-order RNNs by Pollack (1991), Giles et al. (1992), Forcada and Carrasco (1995), Watrous and Kuhn (1992), Ñeco and Forcada (1996), and hybrid or modular RNN architectures by Tiňo and Sajda (1995) and Gori et al. (1998). Others have focused on methods for algorithmically hard-wiring second-order RNNs to behave like a deterministic finite automaton (DFA for short), e.g., Omlin and Giles (1996). The learning capacities of different RNN architectures (including SRNs) are compared in Horne and Giles (1995), and Miller and Giles (1993) have compared the performance of RNNs with the same architecture but first- or second-order excitation function. From this literature, it appears that second-order networks are more suitable for the induction of finite-state automata; learning times and generalization improved when the grammar was complex and these networks were generally faster and more reliable. A common problem witnessed in several studies mentioned above was that RNNs were able to behave like a DFA after training but only for short input strings encountered in training. When tested on longer strings, RNNs were often not able to generalize the learned behavior. Hidden-state representations started to blur for longer dependencies and the RNNs became instable in their DFA emulation. Some of the reasons for this instability have been analyzed by Tiňo et al. (1998) in terms of dynamical systems theory. One strategy to improve generalization is to adjust the network's weights manually after learning. Another option is to extract a formal DFA from the trained RNN dynamics. Such a DFA generalizes perfectly, but the specific mode of network computation is destroyed in this process. Thus, although the RNNs used in the study of regular grammar induction are in principle representationally capable of emulating deterministic finite-state automata, learning such behavior to perfection from limited data is unreliable for currently known methods of RNN training.

2.3.2 Context-free languages

Natural language viewed as the totality of all past, present and future human linguistic behavior is finite and hence trivially regular. Theories of linguistic competence, on the other hand, view language as a productively unbounded system of expressions. Contemporary linguistic wisdom has it that natural language syntax might be accurately modelled by mildly context-sensitive grammars (MCSG). These grammars generate a subclass of the context-sensitive languages which properly includes the context-free languages (CFL). It is widely believed that at least the expressivity of CFLs is required in the formal description of natural language syntax to model the property of clausal self-embedding. Note, however, that self-embedding, which can be described by a rewrite rule $A \rightarrow \alpha A \beta$ with $\alpha, \beta \neq \epsilon$, does not automatically generate non-regular languages (see Chapter 6). Conversely, any CFL can be *effectively* approximated by transforming the CFL into a strongly-regular grammar at the expense of generating a superset of the original language (Mohri and Nederhof, 2001). The use of formal grammars beyond regularity might therefore be justified for reasons of descriptive precision, but also simplicity and parsimony. MCSGs were introduced by Joshi (1995) to model unbounded cross-serial dependencies in Dutch subordinate clauses while at the same time being efficiently parseable.[10] Typologically, such dependencies are rare, CFLs appear to be sufficiently expressive to model the syntax of more than 99.9% of the world's languages. Hence, non-context free languages might be considered necessary mainly to achieve descriptive universality.

The SRN architecture was used in grammar induction by Cleeremans et al. (1989) where it was shown that it could behave like a perfect DFA for simple regular grammars. When the grammar contained long-distance dependencies across embeddings the SRN could still maintain non-local information about dependent elements when the transitional probabilities in the embeddings co-varied with the identity of the head. Their SRN approximated a DFA for short strings and accuracy degenerated with more embedded elements. These results raise the question whether SRN would also be capable of learning nested, non-local dependencies generated by context-free grammars. This question has been addressed most notably by Wiles and colleagues in a series of articles. Like Cleeremans et al. (1989), the learning paradigm they used was the prediction task for sequentially presented input in a backpropagation SRN. Generally speaking, what they found was that the SRN could learn finite fragments of context-free languages without emulating memory devices like counters or stacks. For instance, Wiles and Elman (1995) taught a 2 × 2 × 2 SRN the simplest context-free language $a^n b^n$ in the following sense. The network was trained using BPTT for eight time-steps on a set of 356 sample strings from the language up to depth $n = 11$, strongly biased towards short strings. In testing, the network obviously can not predict the exact length of the initial segment of as. But the network could be considered to perform successfully if, upon the first encountered b, it predicted the remaining number of bs until the end-of-string marker $\#$ was correctly predicted. This criterion subsumed 'acceptance' for formal automata and a rejection response could easily be added to the paradigm. In a population of 20 randomly initialized networks and after being exposed to two million training items, 15 SRN subjects learned the trivial language $a^* b^*$ (any number of as followed by any number of bs, the network parrots its input), four successfully generalized to $n = 12$ and one to $n = 18$. The 'winner' network accomplished this task by developing a damped oscillation dynamics around fixed points in its hidden layer (see Wiles and Elman (1995) for details). Thus, the SRN locally emulated a pushdown automaton but the computational solution did not mimic a pushdown stack.[11] In the authors' words, the network dynamics allowed

[10] In the clause `omdat ik Cecilia Henk de nijlpaarden zag helpen voeren`, for instance, the three verbs depend in serial order on the three consecutive noun phrases (Stabler, 2004).

[11] The approach to connectionist learning of context-free grammars by means of an external stack has

the SRN to 'count without a counter'.

Similar results for learning the language $a^n b^n$ were reported in Rodriguez et al. (1999), Wiles et al. (2001), and Rodriguez (2001) where also the SRN learnability of more complex context-free languages was investigated. The induced solutions reported in these studies, however, appear to be quite unstable and unreliable throughout as they tend to become obliterated again by further training. For example, when training was extended to three million items in the study of Wiles and Elman (1995), many network subjects which had generalized earlier converged on $a^* b^*$ behavior and no subject generalized to $n > 11$. The analysis of Bodén et al. (1999) suggests that this is because the hidden units' solution space lies close to a bifurcation point which causes fluctuating network behavior when trained using any gradient-based weight adaptation.[12] Despite these shortcomings in learning and generalization, it is clear that the SRN of Wiles and Elman (1995) did not merely internalize the statistical properties of the corpus of input patterns. Rather, it developed a genuine computational mechanism to solve the task of learning a context-free language. Using dynamical systems analysis Rodriguez et al. (1999) argued that the induced mechanism was computationally adequate to process strings of unlimited length. The counting dynamics could in principle assume an infinite number of states. Insights from this analysis were then used to manually modify the SRN weights towards a more stable solution which lead to generalization for strings of depth $n = 28$ in some trained models. Hence, it seems that SRNs are potentially capable of learning and representing some non-regular languages to arbitrary depth from finite data.

2.3.3 Context-sensitive languages

Steijvers and Grünwald (1996) investigated whether RNNs could even induce context-sensitive languages (CSL), specifically the simple CSL $(ba^n)^m$. They handcrafted the weights of a second-order RNN which was able to perform a prediction task for strings from this language. The network they developed solved the task for $0 < n \leq 120$ showing that some RNNs are capable of representing some CSLs to considerable depth. The authors did not address the learning problem for their model, i.e., how suitable weights could be induced from exposure to sample strings from this language. They speculate that in general RNNs may not be capable of representing and consequently learning arbitrary CSLs.

The learning aspect for CSLs in RNNs has been studied by Bodén and Wiles (2000, 2002) using the second-order sequential cascaded networks (SCN) first described in Pollack (1991). Training with BPTT showed that SCNs could indeed learn to accurately predict sequences from the CSL $a^n b^n c^n$. From experiencing strings of length $n \leq 10$ during training, their model generalized to $n = 18$ in testing. They also attempted to

been explored in Sun et al. (1990) and Das et al. (1992).

[12] A study of Tonkes et al. (1998) suggested that the seemingly erratic BPTT search for a solution could be improved—in terms of stability and depth, not necessarily speed—when learning was guided by an evolutionary hill-climbing algorithm.

elicit similar performance using an SRN trained with BPTT but failed. Chalup and Blair (1999, 2003), however, employed an SRN together with an incremental hill-climbing algorithm in learning the same language $a^n b^n c^n$, and found good performance at least on the training set; the generalization capacities were not tested.

Returning to the CSL $(ba^n)^m$ of Steijvers and Grünwald (1996), Rodriguez (2001) reported on training an SRN with three hidden units using BPTT for 12 time steps on length-balanced items with $n \leq 7$, $m \leq 10$ for 1.29 million sweeps through the training set. Out of 50 networks, the winning subject managed to generalize all cases of $(ba^n)^m$ with $n \leq 7$ to $m > 10$. The difficulty in processing this CSL, which is beyond the means of a single stack, lies in retaining the 'a-count' across subsequences. Therefore m would seem to be the critical parameter to test generalization on, although m itself is not predictable by the network (as opposed to n). Once the SRN has been exposed to the initial segment $ba \ldots a$ it should be able to predict the number of as when encountering the next b. Unfortunately, Rodriguez did not mention for which m correct prediction broke down (if at all), which makes it difficult to assess the generalization capabilities of the SRN in this task. Furthermore, the network did not generalize to instances of strings with $n > 7$ suggesting that the induced a-counter is not fully general. Yet, as in case of the Wiles and Elman (1995) study, the developed mechanism is computationally adequate and non-contingent as Rodriguez's analysis of the network dynamics shows. The first hidden unit is counting up as, simultaneously the second is counting as down, and the third unit is reactivating the count of unit one for unit two after each segment $ba \ldots a$. Again, an idealized solution can be constructed which expands the discovered dynamics to novel stimuli.

Grammar induction for string languages is an inadequate model of child language acquisition because the primary target of a human learner is to be able to comprehend and produce meaningful sentences. Nonetheless, it is a question of considerable theoretical importance whether RNN architectures such as the SRN are capable of learning string languages from a reasonable amount of positive examples using standard training techniques. The reason is twofold. Demonstrations that they are suitable for this task provide a 'proof of concept' that neural network models of language processing are able to cope with the structural complexity of natural language emphasized in the Chomskyan tradition. And secondly, successful grammar induction shows that connectionist architectures might be capable of developing and generalizing knowledge about purely structural relationships between constituents despite lacking structured representations in the sense of Fodor and Pylyshyn (1988).

In the reviewed experiments it was shown that first-order RNNs could learn to uniformly behave like a deterministic finite automaton when trained in a word prediction task, and they could learn finite fragments of context-free and context-sensitive languages. These languages were not learned in a strict, automata-theoretic sense. Consequently, the learning and generalization capabilities of RNNs remain controversial and uncertain at this moment. Further simulations are required to shed light on the relation between formal grammars and learnable dynamics which is not fully understood. To all appearances, the failure to reliably learn non-regular languages must be attributed mainly to the limitations of the gradient-descent training regime, not the network architecture itself. The widespread conviction that RNNs are at most FSMs is false with certainty in terms of representational competence and false in all probability in terms of learning performance.

Remark on backpropagation

Many of the simulations mentioned in the previous section used variants of backpropagation of error for adjusting the weights in the networks. Severe criticism has been levelled against the neural plausibility of backpropagation learning (Crick, 1989; Zipser and Andersen, 1988). Arguably, it is even contradicting neurobiological facts. Most importantly, it has been criticized that in backpropagation learning error terms are not used locally because there are distinct forward and backward transmission phases. In addition, it is unclear where the error signal originates from in the first place. For grammar induction from string languages these points of criticism might not be relevant since this paradigm lacks psycholinguistic plausibility anyway. However, backpropagation will also be used in all the network models of complex sentence structure I will discuss in the remainder of this chapter, and in the Dual-path model which is the experimental platform in all subsequent chapters. These approaches to syntactic development claim to be more adequate psycholinguistically and attempt to match model performance against human data. Thus, it might be worthwhile to briefly try and defend the use of backpropagation here. Neurocomputational realism is easy to request but difficult to deliver because our limited knowledge of cortical neurobiology does not supply sufficient constraints. Primarily, connectionist models are intended to simulate intelligent behavior not neural activity. The cognitive plausibility of such models is therefore a more important *desideratum* than their neurobiological plausibility. Even though backpropagation may not be a plausible account of learning in the human nervous system, the resulting networks may nonetheless perform a task in a cognitively adequate way once they are trained with it. Secondly, a backpropagation learning trajectory may still accurately reflect progressive stages in human cognitive development. Hence, backpropagation learning may be functionally adequate to produce cognitively and developmentally realistic models, despite being neurobiologically implausible as a mechanism for synaptic change. This is essentially Smolensky's point, that the 'proper treatment of connectionism' lies at the behavioral, not the neural level of analysis (Smolensky, 1988).

Particularly in the domain of language acquisition, backpropagation learning must also be criticized for being ecologically implausible. Language learning is more adequately conceived of as reinforcement learning rather than supervised learning with explicit error signals. Yet, in very general terms, language learning is a task which can be described as the effort to minimize the difference between a learning target and the learner's behavior, and backpropagation is very suitable for modelling this process. Consequently, similar to the responses given above, backpropagation can be justified as being useful functionally and adequate extensionally despite being ill-conceived intensionally (i.e., algorithmically). Nonetheless, it would certainly be desirable to re-place backpropagation by more realistic accounts of synaptic adjustment in biological networks in the long run. Several attempts to approximate backpropagation learning by neurobiologically and ecologically more appealing mechanisms have been made in the literature, e.g., through contrastive Hebbian learning (Xie and Seung, 2003), reinforcement learning (Mazzoni et al., 1991), or recirculation learning (Hinton and McClelland, 1988, and further developed by O'Reilly, 1996). Most recently, Grüning (2007) argued that backpropagation in SRN sequence prediction could be replaced by a reinforcement variant, obtaining similar results in simulations of string language learning. These approaches suggest that the functionality of backpropagation-driven cognitive models might be achievable in multiple ways.

2.4 RNNs and complex sentence structure

Most theories of syntax view natural language as a system of expressions with constituent structure. Constituents are words or groups of words, such as phrases or clauses, which function as a unit within a hierarchical sentence structure. On this view, the syntax of a language specifies the 'legal' relations between constituents within phrases and clauses, and the type of productive patterns to compose hierarchical structure from such units. For instance in English the phrase structure rules

(i) NP → (det) N (PP)
(ii) PP → prep NP

capture the intuition that noun phrases can be prepositionally modified indefinitely as in

(1) `The laptop on the table by the bookshelf in the library...`

and so forth. Sentence (1) is often referred to as a right-branching recursive structure because in order to instantiate either phrasal form on the right-hand side of each rule, a call to the other rule is possible in forming grammatical sentences.[13] Constituency and recursion have been acknowledged as two of the most fundamental linguistic notions even by connectionists (Christiansen and Chater, 2003).

Constituency and recursion are properties of phrase structure rules for generating sentences, but both syntactic properties are not visible in the surface form. Consider the simple string language $a^n b^n$ from above. It could have been generated by the recursive rule of 'self-embedding'

(iii) S → aSb
(iv) S → ab

but it also could have been generated by the instruction 'write any finite number of as

[13] Note, however, that (i) + (ii) are not an instance of true recursion in which a rule would be called by itself. Rather, it is a form of iteration, which could be expressed recursively as NP → (det) N (prep NP).

followed by the same number of bs'. Similarly, the fact that according to (iii) + (iv) the nth a and the first b belong to the same 'clausal constituent' is not encoded in the surface form of the strings in this language. An SRN trained on $a^n b^n$ (or a pushdown automaton, for this matter) could recognize this language without having to recover the constituent structure or the rules of grammar which generated this language in the first place. Both models of language processing, the network and the automaton, need not explicitly represent constituency and recursion but functionally behave as if they did.

In natural language processing, on the other hand, it is widely held that language users represent the constituent structure of utterances and the syntactic rules of generation which give rise to this structure. In comprehension, for example, sentences are presented sequentially to the language processor and their constituent structure must be recovered from this data. How this can be achieved by virtue of internalized rules of grammar is a core *explanandum* of research in natural language processing, such as parsing in computational linguistics. Many accounts seek to identify mechanisms of structure-sensitive processing which induce structured representations from sequential data. It is a fallacy, however, to assume that the abstract structures employed by descriptive linguistics to capture constituency and recursion (such as syntactic trees and phrase structure rules) determine the internal representations the human language processor forms. Moreover, there is no evidence from language processing or acquisition which suggests that causally efficacious rules of grammar are mentally encoded as a stored program which computes representations of sentences (Stabler, 1983). In the course of learning a language processor might become hard-wired to functionally behave in a manner which is consistent with formal rules of grammar without mentally representing such rules in *any* form. Combinatorial syntax is rooted in constituency and generative capacities are rooted in recursive processes. For connectionism it is a central issue to explain how neural network models of natural language processing can come to exhibit behavior which reflects combinatorial syntax and generative capacities without utilizing structured representations and explicit rules of grammar. In this section, I will describe three connectionist models which have attempted to address this issue in some way or another.

2.4.1 Recursive auto-associative memory

Pollack (1988, 1990) devised a neural network model of sentence processing called Recursive Auto-Associative Memory (RAAM for short) which provides a straightforward refutation of the allegation put forth in Fodor and Pylyshyn (1988) that connectionist systems cannot develop representations of combinatorial structure. A RAAM is a three-layer, feed-forward, auto-associative network for encoding and storing linguistic representations of the constituent structure of complex sentences, such as trees and lists, through iterated compression of fixed-width patterns.

Suppose a tree with arbitrary, finite depth d has maximal valence k and it takes at most n bits to represent all nodes of the tree, including the terminal symbols. Then a $kn \times n \times kn$ RAAM exists which encodes such a tree using the following procedure.

First, all terminals $(1_d, \ldots, k_d)$ are presented to the RAAM at its input layer and the network is trained with standard backpropagation to correctly reproduce them at its output layer. This process is called auto-association. Thereby, the n hidden units form a compressed representation CR_d of the pattern of terminals. Moving one level up the tree, the procedure is repeated by auto-associating the pattern of terminals and non-terminals $(1_{d-1}, \ldots, (k-1)_{d-1}, \mathsf{CR}_d)$, and so on iteratively all the way up to the root.[14] In this manner, the RAAM model concurrently evolves two subnetworks during training, one for representing and one for recovering complex tree structures. Once trained on a tree (or a set of trees), the lower part of the RAAM can function as an encoder, whereas the upper part of the network can function as a decoder.

A concrete example may clarify this process. Figure 2.3 shows a simplified phrase structure tree for a sentence with multiple embeddings. Initially the subtree with terminals `Paul loves books` is presented to the RAAM which learns to recreate the input activation pattern on its output. The hidden unit representation of this subtree, denoted `[P/l/b]`, is fed back to the right-most n-width input cluster (Figure 2.4). The network then learns to auto-associate the pattern for `Mary knew [P/l/b]` and so forth until the whole tree is internally compressed. Upon completion of training, the RAAM has developed a 'recursive distributed representation' of the tree at its hidden layer (RDR for short), which can be systematically recovered, reversing the encoding. Repeatedly feeding such a representation into the hidden units will produce an expansion of the tree by one degree at the output layer, until the representation is completely decoded back into the original sentence. This is illustrated in Figure 2.5 (b). Recovery, however, can not

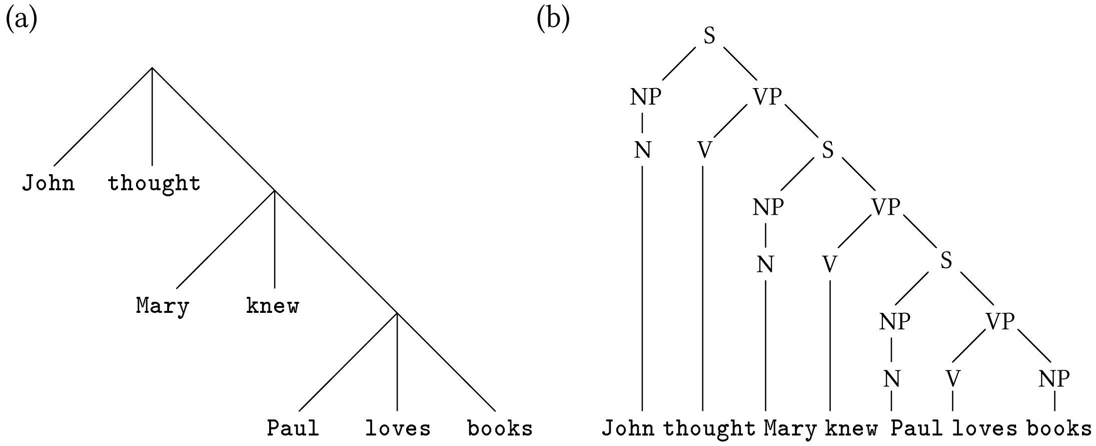


Figure 2.3: (a) A simple representation of the sentence `John thought Mary knew Paul loved books` derived from (b) its phrase structure tree.

[14] This applies only to trees which branch at most once at each level and non-existing nodes are represented by some null-string. The training procedure can be extended to arbitrary trees. First the tree is turned into a full k-ary tree (where k is the maximal valence) by inserting 'empty' nodes. Then all subtrees are encoded serially at each level.

be perfectly accurate because the output activation function is a continuously-valued sigmoid.

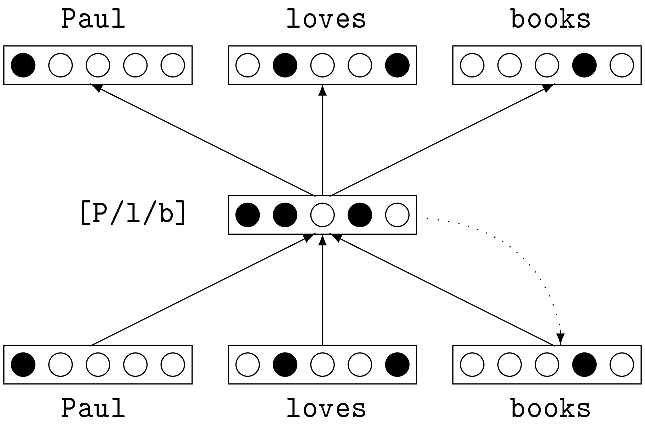


Figure 2.4: A RAAM trained on the terminal subtree of Figure 2.3 (a). Each word is represented at one input cluster and the model recreates the input pattern at its output. The pattern is encoded at the hidden layer. On the next training step this compressed representation becomes part of the subsequent input pattern `Mary knew [P/l/b]` (dotted arrow).

Iterated encoding can accumulate error which causes the decoder to loose track of the constituent structure to such an extent that it mistakes terminals for non-terminals. Furthermore, the unsupervised hidden layer representations, which become partial inputs on the next training step, dynamically change from epoch to epoch, so that the auto-associator is 'chasing a moving target' during encoding, and this variation can not entirely be counterbalanced by the error correction algorithm.[15] Pollack (1990) trained a 48 × 16 × 48 RAAM on a set of thirty phrase structure trees with up to four embeddings

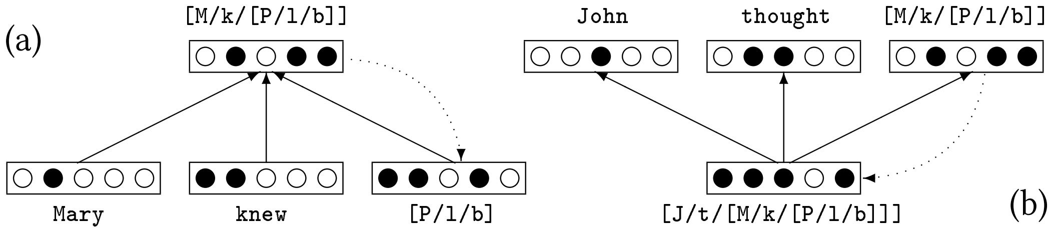


Figure 2.5: (a) After training the auto-associator, the lower part of the network functions as an *encoder*. (b) The upper part functions as a *decoder*.

[15] Consequently, criteria must be imposed to decide when a representation is considered decompressed to a sequence of terminals.

from a simple language. This language consisted of a "somewhat random collection" of meaningful English sentences composed from five word classes, THING, HUMAN, PREP, ADJ and VERB. Despite the shortcomings mentioned above, Pollack was able to show that the decoder of this RAAM could correctly reproduce all trees experienced in training. Because of this capacity of RAAMs to retrieve (an approximation of) the full sentence from a compressed representation, thereby recovering its constituent structure, these models are functionally combinatorial.

Pollack also demonstrated that RAAMs have some structure-sensitive, albeit limited, generalization capabilities. The RAAM in the above experiment was not able to correctly decode representations of novel embedded structures but instead often converged on sentences encountered during training which were within minimal Hamming distance from the target. However, the network correctly recovered simple clause structures such as the 16 instances of the transitive scheme '`X loves Y`' over the lexical items {`John, Mary, Pat, man`}, even though most of these sentences were not in the training set. In other words, the RAAM displayed some systematicity (as characterized in Fodor and Pylyshyn, 1988).

RAAM **applications, extensions, and limitations**

A RAAM generates distributed, fixed-width, real-valued vectors which implicitly encode combinatorial structure. Thus, the representations of a RAAM are essentially of the same data type as the constituents from which they were composed of. It is therefore interesting to investigate whether RAAMs can perform tasks which require sensitivity to constituent structure by operating on such representations *holistically*, without first decoding them into a sentence form.

This issue was taken up by Chalmers (1990) in an active/passive sentence transformation task. First, a RAAM was trained on a corpus of both active and passive sentences to produce recursive distributed representations. Then a three-layer feed-forward transformation network was trained to map an RDR of an active sentence onto an RDR for the corresponding passivized form. To test the tranformation network, an active sample sentence was encoded using the RAAM. The resulting RDR was fed to the transformation network to obtain a passive RDR. This RDR was then decoded back by the RAAM into a passive sentence. Chalmers found that the transformation network generalized perfectly to novel active/passive sentence pairs when decoding errors in the RAAM were systematically eliminated. Blank et al. (1992) investigated similar transformation tasks using RAAMs and Chrisman (1991) extended the approach to holistic computation on RDRs for natural language translation. These studies indicate that neural network models like the RAAM are capable of developing sensitivity to implicitly encoded sentence structure without having to extract and then recombine constituents from the input data.[16]

[16] Chrisman (1991) called this feat 'holistic inference' which maps "directly from the representations of a problem to the representations of its answer in a Gestalt fashion" p. 364.

Two shortcomings of the RAAM architecture appear to impose limitations on its use as a model of natural language processing. First, the RAAM can only deal with fixed-valence trees, and secondly the fixed-size compression layer restricts sentence length and depth of embedding. The latter limitation is due to the hidden units' coarse resolution imposed by finite-precision calculations. In other words, the hidden layer can not encode an unbounded amount of information in any concrete implementation. As Melnik et al. (2000) have argued this is not an architectural constraint. Moreover, it will lead the RAAM to display behavior which degenerates with the number of constituents in a sentence, a trait which may still adequately reflect human performance. The former restriction to fixed-valence trees can be overcome in the sequential recursive auto-associative memory, or SRAAM. Prior to training, tree structures are transformed into sequential lists by encoding internal tree nodes. A complex sentence can then be presented to the network word by word. The hidden layer of an SRAAM functions like a stack whose content is copied back to the input layer at each point in processing. This variation of the RAAM architecture closely resembles an SRN and can be trained in a similar fashion (see Kwasny and Kalman (1995) for details). A combination of recurrent networks and the RAAM model has also been used in the connectionist parser XERIC developed by Berg (1992). This system represents sentence structure by learning to assign syntactic roles of X-bar grammar, such as `specifier`, `head` and `complement`, to constituents. The roles of specifiers and complements can be filled by X-bar structures themselves, hence the XERIC system can in principle cope with embeddings of arbitrary depth.[17]

The RAAM model and all discussed variations thereof reflect top-down approaches to language processing. A network is evolved specifically to represent syntactic relations between locally encoded constituents. Whereas this may prove a useful avenue to modelling human working memory it might not be a viable approach to modelling aspects of learning in connectionist natural language processing. Because of the iterated training RAAMs experience, they do not by themselves induce hierarchical relations from temporal sequences but rather require a complex external control structure which not only stores intermediary representations but *ex ante* constructs a parse tree for the particular training sample. This syntactic preprocessing of the linguistic input is neither removed in the SRAAM model, which operates on sequentially coded trees, nor in Berg's parser where X-bar roles are prompted on the output. The RAAM architecture is therefore primarily a model for the controlled compression of structured data, not for autonomous structure-sensitive learning.

2.4.2 A subsymbolic parser for embedded clauses

In a series of papers, Miikkulainen developed a comprehension model for sentences with relative clauses (Miikkulainen, 1990, 1996, 1997; Miikkulainen and Dyer, 1991; Mi-

[17] In addition, the network was able to perform simple lexical disambiguation regarding word category and number/person.

ikkulainen and Mayberry, 1999). It has a modular architecture which combines an SRN parser with a RAAM memory, controlled by a higher-level, feed-forward segmentation network. A virtue of this model, which will be described subsequently, lies in its motivation to provide a connectionist explanation of attested psycholinguistic phenomena occurring in human processing of embedded structure.

A basic assumption underlying the SRN parser design is that comprehension involves mapping a word sequence into a semantic interpretation which assigns thematic roles to sentence constituents. The SRN learns to decide which words in an input sequence fill thematic roles such as *agent, action, patient, instrument, location, recipient*, etc. (Fillmore, 1968). Through backpropagation the network is trained to assign the current input to a fixed scheme of thematic-role slots and to propose a tentative interpretation for the entire sentence at the same time, including past input and future expectations.[18] This process is depicted in Figure 2.6. An interesting feature of this model component is that it evolves suitable lexical encodings at the input layer from random initial patterns by itself. Thus the network is able to adapt its word representations to the task at hand and is not biased by *ad hoc* choices of the modeler.[19] This subsymbolic parser has been studied in a variety of comprehension tasks (Miikkulainen, 1990, 1997, Miikkulainen and Mayberry, 1999). For instance, the model was trained on an artificial language generated from simple sentence templates and word categories such as HUMAN, FOOD, ANIMAL, PREDATOR, and HITTER. The network learned to assign thematic roles successfully and generalized role assignment to familiar lexical items in

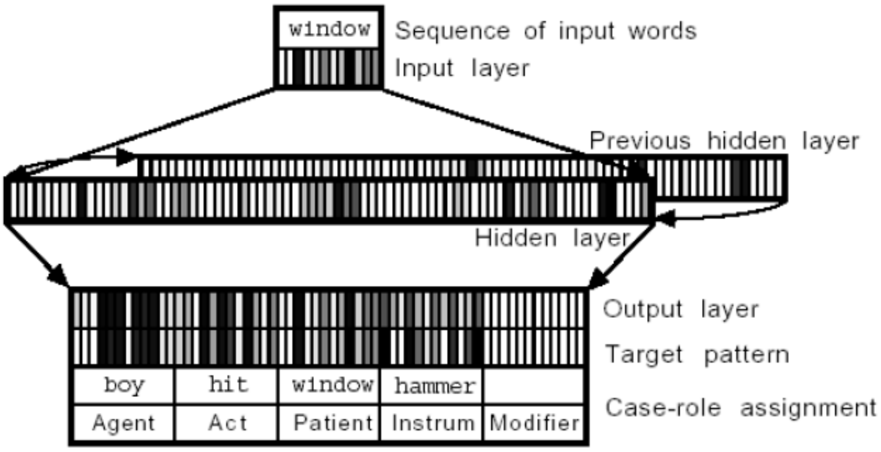


Figure 2.6: The SRN parser performing thematic-role assignment for the sentence `The boy hit the window with a hammer`; image from Miikkulainen (1997).

[18] Thus, different units in the output layer can represent the same word in different thematic roles which is called *binding-by-space*.

[19] This is implemented by extending backpropagation down to the input layer, a training mechanism called FGREP, see Miikkulainen (1990) and Miikkulainen and Dyer (1991).

novel sentences. To achieve this, the model had to learn the semantic regularities in the training set and encode them into the network weights and lexical representations.

The SRN parser also learned to disambiguate lexical meaning depending on a word's context of occurrence, see Miikkulainen and Mayberry (1999). Assigned sentence meaning changed as a function of the lexical meaning of its components and the network was able to tentatively change its interpretation and yet revise it again at a later point in processing if necessary. Thus, the model demonstrated an ability to process sentences incrementally by entertaining the most likely semantic interpretation as it read in one word at a time. In order to process complex sentences with embedded clauses in which multiple thematic roles of the same type could be assigned to different constituents, the SRN had to be augmented with a RAAM dynamic memory. This memory helped the parser to keep track of constituents to which thematic roles were assigned already and which occurred in hierarchical dependencies. The RAAM acted like a stack which stored intermediary representations developed by the parser and popped back previous partial representations into the parser's hidden layer when role assignment in an embedded clause was complete. Accordingly, the target pattern across embeddings changed at each clause boundary so that the parser could effectively learn the correct thematic role assignment as if the input contained no hierarchical structure at all.

An example may clarify this information exchange between modules. Suppose the center-embedded sentence `the girl who liked the dog saw the boy` was received as input structure in the model's role assignment task. The sentence was fed to the SRN parser sequentially, word by word. Initially, the main clause target was a thematic role vector set to t_1=(ACTOR=`girl`, ACTION=`saw`, PATIENT=`boy`). When the relative pronoun was encountered, the parser's hidden layer representation of the incomplete sentence `the girl` was pushed onto the RAAM stack and the parser's context layer was reset. While the subordinate clause was processed, the thematic role target vector switched to t_2=(ACTOR=`girl`, ACTION=`likes`, PATIENT=`dog`). When role assignment was completed, the stored main clause fragment was retrieved from the RAAM and loaded into the parser's context layer. The target pattern then changed back to t_1 for the rest of the sentence. The stack itself worked exactly as described in Section 2.4.1 and there was no *a priori* limit on the depth of embedding that this coupled SRN + RAAM system could process.

It turned out that parser and stack were not yet sufficient for generalizing role assignment to novel relative clause structures. A feed-forward segmentation network was added to the model which not only detected transitions into relative clauses to control the information flow between parser and external memory, but also unified types of relative clauses. This network modified the parser's sequence memory so that, e.g., a center-embedded clause, such as the one in the example above, looked exactly like a center-embedding inside a tail-embedded clause to the parser. The resulting model, which is shown in Figure 2.7, is a full-fledged subsymbolic parser for embedded clauses (SPEC). The division of labor achieved by the modular, tripartite architecture proved very powerful computationally, while at the same time displaying performance limitations that matched characteristics of human sentence processing. The SPEC model

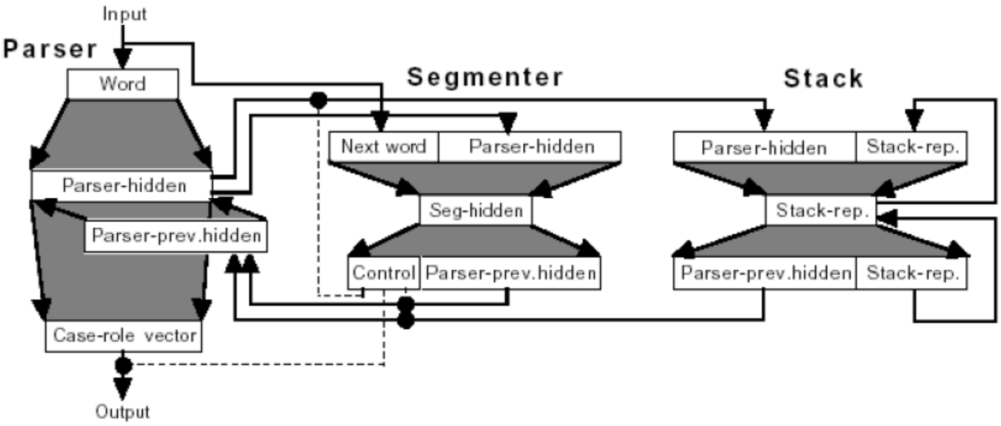


Figure 2.7: SPEC architecture for thematic-role assignment. Grey areas represent total learnable connectivity, solid arrows are copy links, and dotted lines control connections; figure from Miikkulainen and Mayberry (1999).

performed and generalized well in thematic-role assignment tasks for complex sentences with multiple nested embeddings (cf. Miikkulainen, 1996). Trained on a small fraction of the total number of sentences generated by a simple phrase structure grammar, the SPEC model successfully generalized to new instances of familiar sentence templates and also to novel templates. Although in some experiments the SPEC model occasionally confused the roles of two constituents, this virtually always occurred for lexical items of the same word category, e.g., incorrect agent/patient assignment to nouns. This indicates that an abstract categorization of the lexical space in terms of plausible role bindings was acquired.

When subjected to noise, simulating stress and overload, the RAAM component showed degradation which affected the overall system performance as the depth of embedding increased and the stack piled up. The effect closely resembled human error patterns, reported by Miller and Isard (1964) and Foss and Cairns (1970), in that sentences with deeper center-embeddings were more difficult to process and remember than shallow ones. It was also found that semantic constraints between constituents significantly facilitated the task of assigning thematic roles across embeddings, even in the presence of noise.[20] This is in line with psycholinguistic evidence from Stolz (1967) and Huang (1983) showing that human subjects can comprehend relative clause structures much better when semantic constraints are imposed on the relations between constituents.[21]

[20] The strength of semantic constraints was measured, e.g., as the number of nouns that could fill, say, the patient role of a verb. It is not surprising that such constraints influence a neural network model in a sequential prediction task because they immediately affect conditional probabilities between adjacent constituents.

[21] Compare the comprehensibility of the following two sentences (Miikkulainen, 1996):
(i) `The girl who the boy who the man who lived next door blamed hit cried.`
(ii) `The car that the man who the dog that had rabies bit drives is in the garage.`
In sentence (i), every noun could be the subject/object of every verb, whereas in sentence (ii) the argument structure is semantically constrained and can therefore be recovered more easily (e.g., usually it is dogs

The plausibility of SPEC

The behavior of Miikkulainen's subsymbolic parser in the comprehension task is consistent with observations from a number of psycholinguistic studies. But how plausible are the assumptions behind the model's architecture from which this behavior results? First, it should be pointed out that the capacity of SPEC to process complex sentences is largely due to a central executive, the segmenter network, which monitors and controls the combined system. The segmenter detects clause boundaries in the input sequence, recognizes the completeness of a thematic-role vector and the end of a sentence, modifies the parser's context layer, initiates the push/pop operations in short-term memory, and coordinates the temporal dependencies in the communication between modules. Thus, most of the structural demands arising in the parsing process are relegated to a specialized module. Incorporating the segmenter network imposes a modular hierarchy onto the model, but centralized control mechanisms are a feature which the PDP approach to cognition decidedly set out to dispose of. Although the segmenter is implemented by a neural network, the way it performs its task is not distinctly connectionist. It might just as well be realized by a symbolic supervisor.

Secondly, all subnetworks within SPEC are trained separately, without communication between modules. Consequently the learning behavior of each component is not influenced or informed by activity elsewhere in the system. Guided by a dedicated training signal, each module—parser, stack, and segmenter—develops a partial solution to the parsing problem before getting integrated. Cognitive modularity in general is a controversial issue and there is no evidence that language comprehension subdivides into discrete processes developing in complete isolation. Moreover, the segmenter is trained to recognize, not to discover relative clauses. This is a subtle but important difference. The structural knowledge that this control network brings to bear on parsing complex sentences is not itself extracted from sequential data but merely derived from representations of specific clausal segments. Thus, the model's ability to cope with thematic role assignment for multiple embeddings is not strictly learned from temporally extended data but hardwired into the system architecture.

Finally, subnetworks of the SPEC model are trained to perfection and noise has to be added to the memory component to elicit errors which match human performance. This suggests that the performance characteristics of SPEC arise mainly from the properties of the separate memory system, not from the processing mechanism of the parser or the type of representations it employs. Because segmentation and memory are externalized, the parser's task of assigning interpretations to complex sentences is reduced to assigning interpretations to clausal components. Conversely, the semantic processing the parser performs for clausal components does not influence the system's overall performance when decoding complex sentences. Explanations of human linguistic behavior in this model rest on the problematic assumptions that crucial grammatical information and the storage of intermediate representations reside outside the comprehension system. Again, there is nothing distinctly connectionist about such explanations, they

who bite humans, and dogs do not drive cars.).

could be obtained from modular symbolic systems with artificially imposed performance handicaps.

2.4.3 Simple recurrence and complex structure

A more radical approach to processing complex sentences was taken by Elman in a series of experiments (Elman, 1990, 1991, 1993; Weckerly and Elman, 1992). These studies relied exclusively on the processing capabilities of a single, simple-recurrent network without special-purpose modules, external memory or higher-level control. Introducing this particular connectionist paradigm, Elman (1990) trained an SRN to predict the order of words in a linguistic corpus. The corpus consisted of single-clause 2- and 3-word sentences generated from 15 templates over 13 word classes with 29 lexical items which were concatenated into a single input sequence.[22] The network learned to approximate the conditional probabilities of words in a sequence given the previous word by using its short-term memory of prior context. The distributional regularities in the corpus became encoded in the network weights. A cluster analysis of the SRN showed that it had developed representations of lexical categories (noun/verb, animate/inanimate, human/animal, large/small) at its hidden layer from exposure to a continuous input stream. Since the training language lacked recursive structure, no judgement was possible whether the SRN could detect long-distance dependencies as well.

In Elman (1991) an SRN was used to predict word sequences from a context-free language containing complex multi-clause sentences. These were generated from a lexicon of 24 items—8 nouns, 12 verbs, 2 proper names, a relative pronoun *who*, and an end-of-sentence marker #—by the phrase structure grammar shown in Figure 2.8. This artificial language, albeit simple, shared some interesting properties with natural language, for instance restrictions on verb-argument structure, and number agreement between subjects and verbs created long-distance dependencies across relative clauses. The grammar also included recursive rules in that relative clauses (RC) contained noun phrases (NP) which could be modified by another relative clause, and so forth. The language therefore included embeddings of (potentially) arbitrary depth. Moreover, many sentences from this language could legally terminate with an end-of-sentence marker # at several different positions. Compared to processing single-clause sentences in the Elman (1990) study, these properties could be expected to significantly increase the difficulty of the learning task for the network. The architecture used in this task was a 26 × 70 × 26 SRN with two additional 10 unit compression layers immediately below and above the hidden layer. Lexical items were represented locally at the input layer by switching on exactly one corresponding bit. In other words, all items were pairwise orthogonal and thus the network could not read any categorization off the encoding.

Training consisted of four phases of 5 sweeps through 10.000 sentences during which the amount of complex sentences was increased incrementally from 0% to 75% in

[22] These templates were, for instance, [NOUN-human VERB-intransitive] and [NOUN-animate VERB-transitive NOUN-animate].

S	$\longrightarrow$	NP VP #
NP	$\longrightarrow$	PropN | N | N RC
VP	$\longrightarrow$	V (NP)
RC	$\longrightarrow$	*who* NP VP | *who* VP (NP)
N	$\mapsto$	{*boy*, *girl*, *cat*, *dog*, *boys*, *girls*, *cats*, *dogs*}
PropN	$\mapsto$	{*John*, *Mary*}
V	$\mapsto$	{*chase*, *feed*, *see*, *hear*, *walk*, *live*, *chases*, *feeds*, *sees*, *hears*, *walks*, *lives*}

Additional restrictions:

- number agreement between N and V (within clause, and between head N and subordinate V)
- verb arguments:
  chase, feed → require a direct object
  see, hear → optionally allow a direct object
  walk, live → preclude a direct object

Figure 2.8: The phrase structure grammar used in Elman (1991). Constituents in parentheses are optional.

steps of 25%. This methodology was developed in reaction to the finding that when presented with the whole corpus at once the network failed to learn. Focussing on simple input data first, the network could successively learn to handle more complex structure once knowledge of simpler structures was in place. This incremental training strategy was elaborated on in Elman (1993) where it was labelled "the importance of starting small". After training, the network was tested for generalization on a set of novel sentences generated from the same phrase structure grammar. Because the prediction task was inherently non-deterministic, Elman used an error measure which compared the network's output against the statistical probabilities for the occurrence of a word in a given context in the entire corpus. He found that the network performed quite well, producing an overall error of $\rho = 0.852$.[23] According to Elman, the SRN's performance difference between incremental and non-incremental training confirmed Newport's psycholinguistic "less is more" hypothesis (Newport, 1990), that maturational constraints on cognitive resources help language acquisition because they filter out structural complexity in early learning.[24]

[23] The error measure was the mean cosine of the angle between target and output vectors instead of the more common mean squared error which was unsuitable here. The closer the error is to 1 the better the performance.

[24] Rohde and Plaut (1999), on the other hand, argued that neither staged training nor limited working memory are necessary for incremental learning because neural networks can reliably learn local before

The low overall prediction error becomes more significant when analyzing how the network behaved in response to individual complex sentences such as `boys who Mary chases feed cats` during testing. This sentence contains three non-adjacent dependencies between constituents. After having read in the initial segment `boys who Mary`, the network must predict a verb which takes at least one object, the head of the relative clause. This verb has to agree in number with the relative clause subject (`Mary`). When this verb is predicted, the network must omit the direct object from the surface form because the verb occurs in an object-relativized subordinate clause. Instead, the network must produce the main clause verb next which agrees in number with the head noun. Figure 2.9 (a) and (b) show the actual predictions of the SRN after the segments `boys who Mary` and `boys who Mary chases`, respectively. The network correctly predicted a singular, transitive verb which requires a direct object as the continuation of the subordinate clause (Figure 2.9 (a)). Thus, it was aware that the verb class in this position depended on having encountered an object filler (`boys`) previously. After producing the embedded verb, the network predicted plural verbs from different verb classes as possible continuations (Figure 2.9 (b)). Since at this point the SRN could not predict whether the main clause is transitive or intransitive, this is an appropriate output distribution. It is crucial, however, that the network did not activate the class of nouns here to fill the object slot, which shows that it has learned aspects of verb-argument structure for multi-clause sentences. Moreover, all activated verb forms were plural verbs which agreed with the head noun, indicating that the SRN was sensitive to dependencies across the embedded clause.

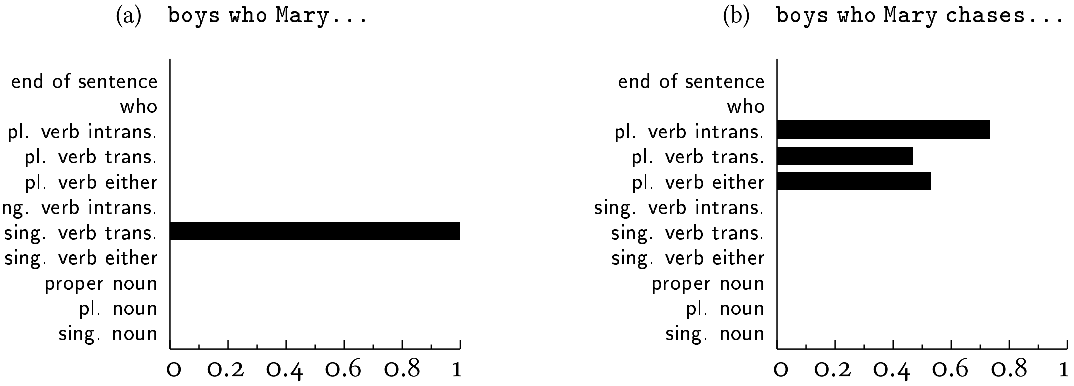


Figure 2.9: Normalized activation vectors in prediction task for a test sentence with relative clause, figures adapted from Elman (1991).

The network's behavior on this and a variety of other test items permitted by the artificial grammar suggests that the SRN (i) developed representations of functional categories (noun/verb), (ii) subcategorized verbs (transitive/intransitive/both) by respecting direct object restrictions, (iii) learned verb-argument structure for complex sentences (presence/absence of direct objects in relative clauses), and (iv) maintained number

non-local dependencies when appropriate semantic constraints are imposed on the language.

agreement across subordinate clauses. Accordingly, Elman argued that connectionist models such as the SRN are computationally adequate to cope with some complex structural relationships between constituents found in natural language.

While these results were an important first step towards modelling constituency and recursion with neural networks, the study of Elman (1991) had a number of shortcomings. The grammar of Figure 2.8 generated subject- and object-relativized subordinate clauses, center-embedded and right-branching constructions, and sentences with arbitrary depth of embedding. It is not clear, however, what Elman's training set looked like in terms of these dimensions of distinction and it was not systematically investigated how the SRN performed on these different structures. Consequently, no attempt was made to compare the model's behavior to human data in this regard. Furthermore, all results were obtained using the controversial staged training regime. And finally, the generalization capabilities of the SRN were not quantified, for instance, by exposing the model to novel sentences with more relative clauses than experienced in learning. It is therefore uncertain whether the model could handle the open-ended productivity of natural language.

2.4.4 Structural processing with semantic constraints

Several of these issues were addressed in a study by Weckerly and Elman (1992). It is experimentally well established that humans display differential behavior in processing distinct types of subordinate clauses. Sentences with center-embeddings such as (2-a), where clauses (NP VP) are infixed between the NP and VP of a higher-order clause twice, are found more difficult to process than semantically equivalent right-branching structures such as (2-b).

(2) a. `The mouse that the cat that the dog scared chased ran away.`
b. `The dog scared the cat that chased the mouse that ran away.`

On the other hand, semantic constraints can significantly improve the comprehensibility of center-embedded sentences (see footnote 21 in Section 2.4.2). Weckerly and Elman (1992) tried to give a connectionist account of why center-embedding is relatively difficult compared with right-branching and how semantic constraints affect structural processing. The SRN used in their simulations was essentially the same as the network in Elman (1991). The lexicon consisted of 10 nouns, 14 verbs, a pronoun `that`, and an end-of-sentence marker.[25] The artificial grammar allowed multiple center-embedding and right-branching and object- as well as subject-relativized constructions.[26] The verb-argument structure of the language is shown in Figure 2.10. As usual, the network was trained in a prediction task using backpropagation learning. Lexical items were coded locally and training was incremental as described in the previous section. The

[25] In generative syntax, `that` is often classified as a complementizer because it occupies the position of true complementizers. I will refer to `that` as a relative pronoun throughout.

[26] Unfortunately, the full generative grammar was not presented in the paper.

Verb	Possible Subject	Possible Object
{walk, live}	HU, AN	—
{write, send}	HU	DOC
{love, kick}	HU	HU, AN
{bite, chase}	AN	HU, AN
{see}	HU, AN	HU, AN, DOC, INANIM
{hear}	HU, AN	HU, AN
{advise, thank}	HU	HU
{own, tame}	HU	AN

Figure 2.10: Argument structure of the artificial language, figure adapted from Weckerly and Elman (1992). Nouns are divided into classes of humans (HU), animals (AN), documents (DOC), and inanimate objects (INANIM).

network was then tested on 192 novel center-embedded and right-branching sentences containing two nested relative clauses.[27] With $\rho = 0.7137$ for the center-embedded versus $\rho = 0.8484$ for the right-branching test set, the network learned to predict the latter construction more reliably. It also showed a degradation in performance with increased depth of embedding similar to human performance data.

Weckerly and Elman offer a genuinely connectionist explanation for these results. In the SRN, grammatical structure is represented at the hidden layer. As the network processes a sentence word-by-word these representations encode the current position in the sentence, the context of the current word, and the possible grammatical trajectories from there. But different types of structures make different representational demands on the processor. Consider a right-branching structure such as (2-b). When reaching the main clause verb `scared` the network can immediately match this verb with the previous, stored noun `dog`. It then expects a suitable direct object after which it can predict the pronoun or the end of the sentence. In the former case, the seen object is expected to be the head of the subordinate clause and upon reading in the next verb `chased` the noun-verb match is already complete. This pattern repeats for each embedding and consequently the network needs to keep track of at most one noun which requires integration with a verb over a short distance. In contrast, the resolution of dependencies in center-embedded constructions is much harder. In order to correctly predict a sentence such as (2-a), the network must store more information over longer a distance. First it reads in three consecutive nouns which need to be kept active in memory simultaneously, then each predicted verb is matched with a subject successively further back in the sentence. At each verb position the network also needs to remember the class of the matching noun (human/animal) to predict an appropriate verb which satisfies the constraints on possible subjects encountered in the training set. These factors complicate

[27] Again, conditional probability distributions were the learning target, and success was measured by the mean cosine ρ of angles between output vectors and empirical likelihoods.

prediction in center-embedded structures compared with right-branching structures and lead to differential behavior. Thus, the performance difference is due to the distance of syntactic dependencies and the memory load the processor has to cope with in center-embedded structures.

Memory load and distance between dependent constituents have also been conjectured to be critical variables in human comprehension (Miller and Chomsky, 1963; Wanner and Maratsos, 1978; Church, 1980; King and Just, 1991; Gibson, 1998). What the model of Weckerly and Elman (1992) suggests, however, is not this 'standard' psycholinguistic explanation of differential performance in which working memory limitations are crucial. Rather, they argue that these limitations themselves arise from the particular representations the processor develops to solve the task at hand:

> "If we view the process of sentence [...] comprehension as movement from one state to another as in a connectionist network, then memory limitations [...] are due to the nature of representations (in human memory) in sentence processing."
> (Weckerly and Elman, 1992, p. 417)

The structural information necessary to predict a word sequence at each point is encoded in a fixed-width state-vector guiding the processor through 'grammatical space'. The amount of data and the temporal distance between dependent data points in center-embedded structures is particularly taxing the network's representational capacities. Information from different embeddings is concurrently active in the network's state-vector and this impedes the prediction of the next word. The processing limitations of the model therefore arise from representational demands within the processor itself and are not a consequence of the limitations of a separate working memory system external to the processor.

The effect of semantic constraints

In a second experiment, Weckerly and Elman (1992) examined the effect of semantically constrained verb classes on the processing of center-embedded constructions. The network was trained as before but tested on two distinct sets of 192 novel sentences. The first set contained only sentences in which each verb admitted a unique class of subjects, objects or both (according to the semantic structure given in Figure 2.10). The second set contained only sentences where subjects and objects belonged to different noun classes. An example sentence from the first set is given by (3-a),

(3) a. `Dog that Dorothy that bear bites tames chases tiger.`
b. `Dog that Dorothy that bear sees hears walks.`

sentence (3-b) is an example from the second set. In (3-a) the subject of `bites`, for example, must be an animal, whereas in (3-b) humans and animals are both possible subjects of `sees`. It was found that for one as well as two levels of embedding the semantically constrained corpus was predicted more accurately than the unconstrained corpus and this observation is consistent with human performance data (Blaubergs and Braine, 1974; Stolz, 1967).

The explanation of this behavior in the SRN is two-fold. When the first verb is encountered in either of the sentences (3-a) and (3-b), the network already 'stores' three nouns serving as potential subjects. But the subject-verb resolution is facilitated in (3-a) by a direct incompatibility with one of the memorized nouns (`Dorothy`); a human subject cannot function as the subject of `bites` in this artificial language. At the same time, because this resolution is facilitated, the network is put in a more distinct state of expectation about the next verb, i.e., the likelihood that the next verb takes the noun `Dorothy` as its subject increases. In this bi-directional fashion, semantic constraints aid in the resolution and prediction of syntactic dependencies.

It is perhaps not particularly surprising that semantic restrictions of this kind should increase prediction accuracy in the model or comprehensibility in humans. What this model demonstrates in addition, however, is that semantic processing and syntactic parsing can interact in parallel. Sequential as well as 'semantic information' are not distinct, encapsulated information types but are available to the processor simultaneously at all levels of embedding.

2.4.5 Increasing grammatical complexity

The focus of Elman's language processing experiments was to show that SRNs can learn aspects of complex grammatical structure from sequential data under favorable conditions such as incremental training. Following up on Elman's work, Christiansen (1994) and Christiansen and Chater (1999b) pushed the processing load of SRNs further by investigating the 'recursive' capacities of this model. The idea was to obtain a more systematic picture of SRN behavior by identifying those neuralgic properties of artificial grammars which elicit prediction failure.

Christiansen (1994) devised two phrase structure grammars which admitted several constructions found in natural language that were not encompassed by the grammars of Elman (1990, 1991, 1993). The first grammar contained the following generative devices:

(i) *left-branching recursion* in the form of iterated genitives (`John's boys' dogs`)

(ii) *right-branching recursion* in the form of prepositional modification (`city near lake`), conjunction (`John and Mary`), sentential complements (`Mary says that John knows`), and subject-relative clauses (`boy chases girl that runs`).

(iii) *mirror recursion* in the form of center-embedded object-relative clauses (`cats who John chases run`).

The full grammar and lexicon are shown in Figure 2.11.

The second grammar replaced center-embedding with cross-serial dependencies.[28] Both grammars generated sentences over the same vocabulary and could express the

[28] The object relative clause construction, which created center-embedding, was removed.

S	→	NP VP #
NP	→	PropN $\mid$ N $\mid$ N rel $\mid$ N PP $\mid$ gen N $\mid$ N *and* NP
VP	→	V(i) $\mid$ V(t) NP $\mid$ V(o) (NP) $\mid$ V(c) *that* S
rel	→	*who* NP VP(t/o) $\mid$ *who* VP
PP	→	prep prepN
gen	→	N + "s" $\mid$ gen N + "s"
N	↦	{*boy*, *girl*, *man*, *boys*, *girls*, *men*, *cats*, *dogs*}
PropN	↦	{*John*, *Mary*}
V(i)	↦	{*runs*, *jumps*, *run*, *jump*}
V(t)	↦	{*loves*, *chases*, *love*, *chase*}
V(o)	↦	{*sees*, *see*}
V(c)	↦	{*thinks*, *says*, *knows*, *think*, *say*, *know*}
prep	↦	{*near*, *from*, *in*}
prepN	↦	{*town*, *lake*, *city*}

Figure 2.11: The phrase structure grammar from Christiansen (1994) which permitted center-embedding. V(i) stands for intransitive verbs, V(t) for transitive verbs, V(o) is optionally transitive, and V(c) are verbs expressing propositional attitudes.

same sentential content with different constructions. Because of the diversity of recursive constructions in these grammars, they imposed computational demands on the SRN which were well beyond those of Elman's studies.

Using a standard 42 × 150 × 42 SRN model in a word prediction task, Christiansen (1994) conducted a number of experiments on learning and generalization with these grammars. Throughout, training and test sets contained 10000 randomly generated sentences of variable length and syntactic complexity. Training was incremental and consisted of 5 phases. Unlike in Elman's studies, in each phase the entire corpus was presented to the network for several sweeps. Maturational constraints were simulated by limiting the hidden layer's 'memory window'. The context layer was periodically reset after n words with n growing across training phases.

At first, the SRN's general performance on sentences of at most one level of embedding was separately evaluated and the network performed very well on both grammars in terms of the mean cosine measure. In addition, the results were markedly above the performance of n-gram statistical models ($1 \leq n \leq 5$), indicating that the network had learned some complex structural regularities, not merely relative word frequencies. In the next simulation, the depth of embedding was increased. Apart from left-branching constructions (prenominal genitives) and right-branching constructions (complement clauses, prepositional modification, conjunction, subject relative clauses) the network was exposed to doubly center-embedded sentences (such as `cats who John who dogs love chases run`) which contained two nested object relative clauses. Both required transitive verbs that took the two initial nouns as their direct objects. Although the network did make partially correct predictions for these structures, it showed a peculiar 'breakdown pattern' (see the histograms of Figure 2.12). The network is on target still in 2.12 (a) where it predicts a plural transitive verb but in 2.12 (b) prediction starts to go awry. The model should exclusively have predicted a singular transitive verb to match `John` but instead activated all other verb classes as well, including all plural forms. Moreover, despite two open subject-verb dependencies the network opted to abort the sentence at this point by activating the end-of-sentence marker. When it received the intended continuation `chases` it again failed to activate another plural form in 2.12 (c) and activated single and plural nouns and the end-of-sentence marker. At this sentence position sentence the network has gone wrong completely but once it sees the verb `run` it recovers to correctly predict sentence termination. The same experiment was conducted with the cross-serial dependency grammar and Christiansen (1994) reported similar findings for sentences with two crossed dependencies such as `dogs John cats love chases run`. In this English paraphrase, `dogs` is the subject of `love`, `John` the subject of `chases` and the object of `love`, and so on. The network struggled particularly with the second verb but the prediction error was not as severe as in the doubly center-embedded case and recovery was better. This indicates, Christiansen pointed out, that the network performed better on cross-serial dependencies than on center-embeddings which is in accord with findings from a study by Bach et al. (1986). Note, however, that the cross-serial word sequence looks to the network exactly like a doubly center-embedded sentence with pronouns omitted. This suggests that the performance difference for the two types of recursion might result from interspersed pronouns in center-embedded sentences, rather than from different kinds of dependencies in the two structures.

Finally, Christiansen (1994) examined the network's performance when processing instances of multiple branching for the recursive constructions permitted by both grammars (i.e., prenominal genitives, complement clauses, right-branching subject-relative clauses, and prepositional modification). The general pattern observed for these structures was that prediction accuracy slowly degraded (with the exception of sentential complements) as recursive depth increased and sentences became more complex. Broadly speaking, due to memory limitations this behavior might be expected in humans as well, although experimental data on this issue is lacking.

Discussion of Christiansen's results

The study of Christiansen (1994) was the first to systematically investigate the processing of recursive structure in SRN and provided important insights into the capabilities and limitations of this model when learning complex grammars which capture many properties of natural language. Christiansen concluded from the experiments on multiple center-embeddings and cross-serial dependencies that the former are harder to process than the latter, although they are computationally more costly in terms of the Chomsky hierarchy. He suggested that the network's behavior matched human performance according to several psycholinguistic studies of recall, comprehension and grammaticality judgment. Moreover, right-branching subject relative clauses appeared

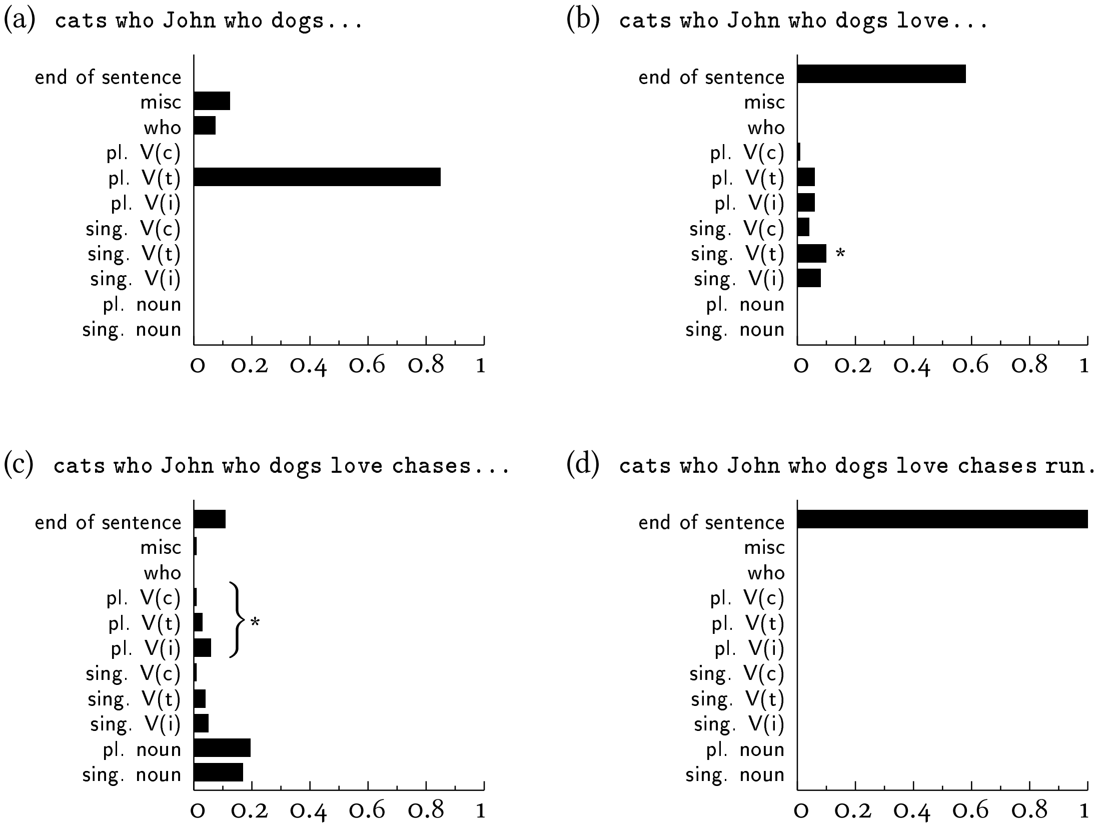


Figure 2.12: SRN predictions for a test sentence with double center-embedding, figures adapted from Christiansen (1994). The category `misc` included `that`, `and`, genitive markers, and prepositions. Intended predictions are marked with an asterisk.

to be easier for the SRN than both center-embeddings and cross-serial dependencies and the network's performance degraded in a fashion similar to humans for most of the recursive structures tested as the depth of embedding increased.

The results summarized in the previous section, are difficult to interpret for a number of reasons. First, the performance of the networks in the testing phase was not rigorously quantified in terms of, e.g., the mean cosine measure. Assertions about the relative difficulty in processing different recursive structures were made on the basis of examining the histograms of single sentences. Without quantitative data for a larger corpus of test items, the network's behavior is not comparable across different tasks and claims about differential behavior are hard to vindicate. Secondly, structures were compared by testing networks which did not experience the same training. For example, Christiansen cites several psycholinguistic studies showing that multiple subject relative clauses are easier to process than multiple object relative clauses. The corresponding experiments in Christiansen (1994) do not strictly warrant this conclusion for the SRN because subject relative clause performance was tested on a network trained with the cross-dependency grammar while object relative clause performance was tested on a network trained with the center-embedded grammar. Third, on a high level of analysis, network and human processing behavior appear congruent in a number of tasks. Erroneous predictions of the network are interpreted as processing difficulties which humans have too. The network's breakdown patterns, however, have no correlate in psycholinguistic data and are unlikely to be observable or reproducible in humans (which Christiansen concedes). These patterns are not pointing towards a specific kind of processing problem in humans. Regrettably, Christiansen did also not attempt to analyze the internal representations of the SRNs to explain differential behavior across recursive types and the nature of failure in each particular task.

Most of these issues have subsequently been addressed in Christiansen and Chater (1999b). In this study, SRNs were exposed to artificial grammars which generated a variety of structures by means of recursive rules such as, among others, right-branching subject relative clauses, center-embedded object relative clauses and relative clauses with cross-serial dependencies. The grammars used in these experiments were not as complex as the phrase structure grammar of Figure 2.12 since the aim was to conduct benchmark tests for recursive types in a simplified and pure form, without the potential influence of other constructions in the language on learning and generalizing recursive structure. In two of the simulations, an SRN was trained on a language which permitted right-branching recursion and either center-embedding or cross-serial dependencies. The input set contained 5000 items of which 30% were single-clause sentences. The complex sentences were split between the two recursive types with various levels of embedding (55% of the total training set with one embedding, 14% with two embeddings and 1% with three embeddings). Each network was tested on 500 novel sentences from the same grammars. The right-branching structures were present in all languages and served as a baseline against which SRN performance on other types of recursion was evaluated. What Christiansen and Chater (1999b) found was that the SRN tested better on cross-serial dependencies than on center-embeddings and these results did not depend on the size of the SRN's hidden layer (between 2 and 100 units were tested). Moreover, the difference in performance did not depend on the amount of training on sentences with deep embeddings. These results indicate that differential behavior did not derive from arbitrary memory limitations in the network but reflected a genuine, intrinsic processing bias of the model. Secondly, they found that generally the SRN performance degraded with the depth of embedding. For 1–4 embeddings the model performed better on center-embeddings and cross-serial dependencies than on right-branching, for 2–4 embeddings it performed better on cross-serial than on center-embedded sentences. In other words, with increased depth the prediction error increased more strongly on center-embedded sentences than on cross-serial dependencies. In contrast, the prediction error for right-branching recursion only increased mildly with depth in the center-embedded grammar and even decreased slightly in the cross-serial dependency grammar. These results provide a good fit with human processing data for multiple embeddings in similar construction (see the discussion in Christiansen and Chater (1999b) for more details). To summarize, Christiansen and Chater have shown that there is a "close qualitative similarity between the breakdown patterns in human and SRN processing when faced with complex recursive structures” (p. 201). Thus, they have given a learning-based, connectionist account of recursion in linguistic performance which does not require imposing *ad hoc* limitations on human grammatical competence with unbounded recursive capacities.

The SRN simulations of Christiansen and Chater (1999b) were conducted by separating different types of recursive structures into several grammars. The network was exposed to each grammar in turn and performance was compared within and across the individual experiments. Although this methodology might be adequate to uncover processing biases inherent to the SRN architecture, it can be criticized when applied to explaining human behavior. A human learner of Dutch, for instance, is exposed to right-branching, center-embedded and cross-serial dependencies during acquisition and must learn all structures concurrently. It would therefore perhaps be more adequate to test the SRN on a grammar which generates all three of these structures. There is reason to suspect, however, that the SRN might not display the RB $<$ CS $<$ CE performance ordering reported in Christiansen and Chater (1999b) when exposed to sentences from such a grammar.[29] This is because the SRN is sensitive to substructure frequencies in its input and the RB $+$ CS $+$ CE grammar might give rise to substructure frequencies which facilitate or encumber the learning of each recursive structure in a different way than the RB $+$ CS or RB $+$ CE grammars in isolation. To test this idea, I performed a simple bigram analysis for three string languages which roughly corresponded to the grammars used in Christiansen and Chater (1999b). In each language, lowercase letters a, b, c denote nouns and uppercase letters A, B, C denote verbs. Nouns and verbs formed strict pairs aA, bB and cC in order to express distinct dependencies in strings. Thus, the RB-language consisted of strings such as $aAbBcC$, $cCaAbB$, and so forth, indicating that surface dependencies were adjacent in each clause. The CE-language consisted of all strings of the form $abcCBA$ where dependencies were mirrored, and the CS-language consisted of all strings $abcABC$, so that dependencies were cross-serial. These languages permitted two and only two levels of embedding. Bigram probabilities were then computed for the strings of the RB $+$ CS and the RB $+$ CE language (12 strings in each language). It was found that the bigram model predicted the order RB $<$ CS for the former and RB $<$ CE for the latter language. Comparing string probabilities across languages showed that CE $<$ CS and RB(CE) $<$ RB(CS), i.e., RB-strings had a lower predictive probability in the CE-language than in the CS-language. Notice that these four orderings are precisely what Christiansen and Chater (1999b) found for sentences with two embeddings from their languages using a bigram model (p. 181). When probabilities were computed for strings of the RB $+$ CS $+$ CE language, however, the order CE $<$ RB $<$ CS was obtained. Strings from the CE-language had the highest probability in the bigram model because substructures such as ab or BC could occur in both CE-strings and CS-strings but not in RB-strings. This substructure over-

[29] I will use the abbreviations RB = right-branching, CS = cross-serial dependencies, and CE = center-embedding in the remainder of this discussion. RB $<$ CS $<$ CE means that RB-structures are easier than CS-structures, which are easier than CE-structures.

lap pushed the string probabilities of CE-strings above those of RB-strings which means they were easier to predict based on exposure to all strings from the RB + CS + CE language. This highly simplified model shows that two 'input' grammars in isolation can predict behavior which is partially consistent with human data (RB < CS and RB < CE) but jointly they predict behavior which mismatches human data (CE < RB). As Christiansen and Chater (1999b) demonstrate, the bigram (and the trigram) model is not an accurate predictor of SRN behavior in a number of respects (e.g., the latter displays CS < CE whereas the former does not). Nonetheless, it is conceivable that substructure frequencies in a RB + CS + CE language might also affect the differential behavior of the SRN on recursion types in undesirable ways and partially reverse the performance order. To put it differently, the SRN might have an intrinsic processing preference for one type of dependency over another (as Christiansen and Chater (1999b) argue) but this bias might be erased by different distributional regularities in the input. In fact, it will be argued in Chapter 8.4 that the Dual-path model (see Chapter 3) can account for a human acquisition and processing hierarchy of different relative clause constructions based on the types of structures in the input, and the Dual-path model includes an SRN as a sequencing submodel. The hierarchy arises solely due to patterns of similarity and interference between structures in the language to which the model is exposed. To be a viable model of recursion in human linguistic performance, the SRN of Christiansen and Chater (1999b) in my view would have to be trained on a language containing all three types of recursion (RB, CS, and CE) to see if the RB < CS < CE processing order persists.

2.5 Summary

I started this chapter by surveying formal results from the study of neural networks as mathematical objects. The results suggested that there is no reason to reject these models on grounds of their computational inadequacy for natural language processing. The widely used class of simple-recurrent networks, for instance, was proven to be computationally equivalent to the class of finite-state machines when implemented with fixed precision arithmetic (Kremer, 1995). When this assumption is dropped, first-order, heterogeneous recurrent networks with rational weights are Turing-equivalent (Siegelmann and Sontag, 1991). Thus, neural networks generally are a very powerful class of computational devices.

The representational capacities of neural networks, however, must be distinguished from their learning capacities. Whether a given network topology, a set of input/output patterns and a training procedure are sufficient to make this network learn the appropriate mappings is certainly amenable to mathematical analysis, but in practice it is mostly treated as an empirical question.[30] I reviewed some of the simulations which

[30] In language processing, the function a network is supposed to learn is often not explicitly specified by the experimenter, but only implicitly through the mechanism which generates a set of input/output patterns on which the model is trained.

address this question in the domain of string language learning, using the simple-recurrent network. It was found that this model could induce recognizers for languages of varying automata-theoretic complexity. When exposed to samples from a simple regular grammar, the SRN could perfectly emulate a deterministic finite automaton. Due to the temporally extended recurrence in SRN, there is no *a priori* reason why these systems should not be able to also track dependencies over a distance. It was shown that SRN can learn long-distance dependencies (under certain conditions) and that the reliability of this process degrades with distance (Cleeremans et al., 1989, Servan-Schreiber et al., 1991). Recently, Onnis (2003) demonstrated that the ability of SRN to predict agreement dependencies is modulated by the variability of the material which intervenes, and this was in line with human data. For non-regular languages, the SRN appeared to be limited to embeddings of small depth and generalization was unstable (Wiles and Elman, 1995, Steijvers and Grünwald, 1996, Rodriguez, 2001). Both limitations may result from the deficiencies of backpropagation learning rather than the SRN architecture itself (Bengio et al., 1994).

I then discussed a number of connectionist models which have been proposed in the literature specifically to deal with aspects of recursive syntax found in natural language. The RAAM model was able to develop compressed, distributed representations of phrase structure trees of complex sentences (Pollack, 1990). These representations encoded constituency in a holistic fashion and could be used in structure sensitive processing. The SPEC model which was built from a RAAM memory, an SRN parser, and a feed-forward sequencing network performed thematic-role assignment in the comprehension of multi-clause utterances (Miikkulainen, 1996). Several pioneering SRN studies demonstrated that connectionist systems are suitable models of learning artificial languages which contained more lexical and structural diversity than string languages (Elman, 1990, 1991, 1993). Christiansen, finally, pushed the limits of the SRN and showed that this model might be able to explain the differential processing of recursive types in humans (Christiansen, 1994; Christiansen and Chater, 1999b).

Chapter 3

The Dual-path model

In this chapter I describe and motivate the architecture of the Dual-path production model and its semantic representations. An example will be given to illustrate how the model produces sentences from such representations. I will then review important properties of this model and summarize past research with it.

3.1 Limitations of the SRN approach

The adequacy of neural networks, and specifically the SRN, as models of language learning and processing has been challenged in a number of ways. But few of these criticisms seem to be pointing towards fundamental limitations of the SRN architecture as such. Rather they are often based on preconceptions regarding the nature of syntactic representations. It might be argued, for instance, that unlike humans the SRN does not 'truly' represent hierarchical phrase structure or long-distance dependencies but merely records transitional probabilities between constituents. Objections along these lines are irrefutable by computational simulations, because no behaviorally adequate connectionist system could invalidate the empirical premiss of the argument. A good example of this kind of controversy can be found in the debate ensuing a study by Marcus et al. (1999) of rule-learning in infants (Altmann and Dienes, 1999; Christiansen and Curtin, 1999b,a; Eimas, 1999; Marcus, 1999a,b,c; McClelland and Plaut, 1999; Negishi, 1999; Seidenberg and Elman, 1999a,b). Nonetheless, the standard SRN model is limited when compared to human linguistic behavior. These limitations, however, do not invalidate the SRN approach in a principled manner. Moreover, they can be overcome by extending the SRN architecture appropriately and this chapter describes an attempt in this direction.

3.1.1 Meaning

Learning a language involves learning to map meaning representations (message for short) onto sequences of words in production, and vice versa in comprehension. This mapping is mediated and constrained by syntactic knowledge. Thus, linguistic behavior is a transduction process between different kinds of representations, messages and sequences of words. Standard SRNs which have been used in language learning (e.g., Servan-Schreiber et al., 1991; Elman, 1991; Christiansen and Chater, 1999b) map sequences of words onto sequences of word categories as shown in Figure 3.1. Perhaps the most obvious limitation of these models is that they do not explicitly represent lexical or sentence meaning. When processing sentences, the SRN does not associate semantic representations with word sequences and is therefore neither a model of production or comprehension proper. Rather, an SRN can be viewed as a stochastic part-of speech tagger which predicts grammatical sequences of word categories (see Steedman, 1999). Such stochastic tagging may be part of the human language system, in particular in comprehension, but it is not an overt process in human linguistic behavior. Of course, the lack of meaning in standard SRN is not an inherent limitation of the mechanism. In order to function as either comprehension or production model, the SRN can be augmented with a meaning system (e.g., St. John and McClelland, 1992; Dell et al., 1999). In these models, the SRN is a subcomponent which helps to learn mappings between word sequences and semantic features. The Dual-path model (Chang, 2002; Chang, Dell, and Bock, 2006) which is introduced in this chapter is such an extension of the SRN architecture which can both represent the intended meaning of a sequence of words as well as the semantics of words in its lexical layers.

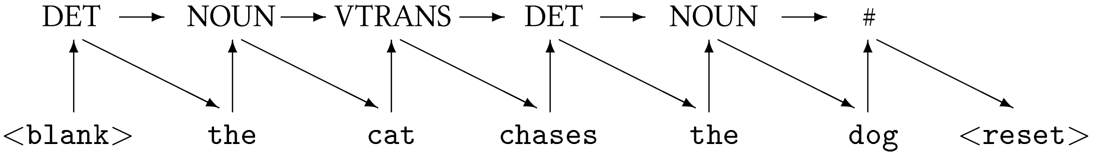


Figure 3.1: Sequential computation of an SRN in a word prediction task.

Adding semantic representations to the SRN is not merely a gimmick to render the model more realistic with respect to human linguistic behavior. The use of semantic information might be an essential aspect of acquisition and processing itself. Adult speakers use language to convey meaning, and it has been argued that children must also use meaning in syntactic development if they are to acquire adult-like linguistic representations (MacNamara, 1972; Pinker, 1984; Tomasello, 2003). In the development of vocabulary, for instance, it has been suggested that children draw on observation, world knowledge and other extra-linguistic context to first establish the semantic properties of words. From these properties, word categories are inferred and the syntactic relations between words can be derived from the semantic relations between referents in observed events. On this view of syntactic development (*semantic bootstrapping*), "semantic representation[s] [are] part of the input to the language acquisition mechanisms" (Pinker, 1989, p. 39). On a different view (*syntactic bootstrapping*), syntax is a cue to word meaning in that, for example, children supposedly rely on the number and types of arguments of a verb to infer its meaning (Fisher et al., 1991). In early linguistic experience, children often simultaneously observe objects and events while listening to adult speech. The visual stimulus and the syntactic context of word occurrence have to be paired. The semantic and syntactic bootstrapping hypotheses differ in their direction of explanation. On the former view, meaning is inferred from the visual stimulus and drives syntactic development, on the latter syntactic context drives the construction of meaning which is mapped onto the visual stimulus. The vast amount of literature and the complexity of the issue prohibit a closer look at this debate here. It appears, though, that the controversy is difficult to resolve for a number of reasons. First, different processes might drive the acquisition of different classes of words. Semantic bootstrapping might better explain noun learning because referents are usually objects in the visual environment, while syntactic bootstrapping might better explain verb learning. Notice also that much of the experimental evidence for syntactic bootstrapping of verb meaning with nonce words depends on previously learned nominals occurring in a syntactic frame. Secondly, both kinds of bootstrapping might operate in parallel. A child might infer the meaning of a word from the visual stimulus and then check whether this interpretation is consistent with the syntactic context in which the word occurred. Third, it is difficult to nail either hypothesis experimentally, because often an alternative account of the data can be given from the opposite theoretical angle. In a classical study by Brown (1957), for instance, children were presented with sentences such as *He's daxing him* together with pictures of unusual events and they chose a picture of an action as the referent of *daxing*, rather than an object or substance. It was suggested that it is the verb morphology (-ing) which supported this choice. But it might just as well be argued that children inferred the thematic core of 'someone doing something to someone else' and chose an action because they could fix the referents of the pronouns in the depicted event. In this way, the meaning of the novel word *daxing* could have been constructed from the meaning of the utterance as a whole.

A related point has been stressed in the acquisition of argument structure (Gropen et al., 1989). Purely syntactic accounts cannot easily explain why some verbs can occur in the dative alternation, but others can only occur in prepositional datives (but not the double object dative). If events are construed as 'causing a thing to change location' (prepositional dative) versus 'causing a person to change his possessions' (double object dative), however, not all verb meanings are compatible with change of possession in a double object frame (e.g., drive, push). Thus, the meaning of verbs is not determined by syntactic variations in the phrase structure in which they occur (syntactic bootstrapping), but on the contrary the syntactic properties of verbs (in which dative structure they could be used) derive from the semantics of the entire construction.[1] On this account, meaning is primary not derivative, it is the *explanans* of syntactic development rather than its outcome. In order to capture the idea that meaning drives syntactic development, a computational learning model needs to be equipped with semantic representations.

In adult processing a similar divide exists between syntacto-centric and multiple-constraints accounts which emphasize the role of meaning and other sources of infor-

[1] See also Tomasello (2003), Goldberg (2006), and Chapter 5.

mation in the resolution of ambiguities during comprehension. The view that syntactic processing is prior to and largely independent of semantic and contextual processing has been advocated, e.g., by Frazier and Rayner (1982) and Ferreira and Clifton (1986). On these accounts ambiguities are resolved first in a purely syntactic fashion, regardless of available non-syntactic information. Multiple-constraints accounts, on the other hand, propose that the processor continuously evaluates several syntactic alternatives against information from the non-linguistic context and takes into consideration the semantic and thematic appropriateness of possible arguments in each phrase (MacDonald et al., 1994; Trueswell and Tanenhaus, 1994; Spivey-Knowlton and Sedivy, 1995). These accounts are supported by a wealth of more recent evidence from psycholinguistic and neuroscientific research, suggesting that non-syntactic information directly influences comprehension. In eye-tracking studies of comprehension, for instance, it has been shown that visual information and knowledge about events and their participants guides the interpretation of sentences before syntactic dependencies (between verb and direct object in single-clause sentences) are resolved (Altmann and Kamide, 1999; Kamide et al., 2003). The influence of non-linguistic context has also been demonstrated in the resolution of syntactic ambiguities created by prepositional phrases (Spivey et al., 2002). Semantic effects on comprehension have been investigated in a number of studies and it was shown that the thematic fit between sentence-initial nouns and a particular verb can influence expectations in comprehension (Trueswell et al., 1994; McRae et al., 1997; McRae et al., 1998; Spivey-Knowlton and Tanenhaus, 1998). Ferreira (2003) has argued that semantic anomalies can determine the interpretation of sentences even when this interpretation is in conflict with the syntactically unambiguous surface form. Recording event-related potentials during sentence comprehension, several studies indicate that semantic and syntactic processing operate in parallel (Kim and Osterhout, 2005) and are integrated as soon as relevant information becomes available (Brink and Hagoort, 2004; Friedrich and Kotz, 2007). These findings suggest that meaning and other types of non-syntactic information are non-redundant in language learning and processing, and point to a fundamental limitation of the SRN approach.

3.1.2 Definiteness

Natural language syntax includes constructions such as relative clauses which separate the head noun from the verb that agrees with it in number. For instance, in the sentence

(1) `The boys that the dog chased are playing in the garden.`

the head noun `the boys` requires the plural auxiliary `are` after the relative clause. This relation between elements is often called unbounded long-distance dependency because there is no *a priori* bound on how many embeddings can separate dependent constituents. It is crucial to the processing of natural language syntax that such non-adjacent dependencies can be maintained. Although SRN learning is based on extracting dependencies between adjacent constituents, it was argued in the previous chapter that under some conditions SRN can detect non-adjacent dependencies due to time-delayed recurrence. Once the relative clause in (1) is complete, the prediction task of the SRN becomes deterministic in that a verb form must be predicted which agrees with the head noun in number. Predictions of the structural properties of the content between the dependent elements, however, are non-deterministic. The SRN activates every grammatical continuation which is consistent with the input distribution at choice points between non-adjacent dependencies. Consider a simple language which only contains an article, singular nouns, a relative pronoun and transitive verbs, and which can express sentences such as (2) and various types of relative clause constructions such as (3).

(2) `The boy hit the man.`

(3) `The boy who the dog chased hit the man.`

When presented with the joint initial segment of (2) and (3), the noun phrase `the boy`, an SRN trained on such a language, predicts the onset of the embedded clause in the weak sense of activating lexical items from several word categories which are possible legal continuations. For example, it might activate the class of transitive verbs and also the relative pronoun, depending on the frequency of subject-modifying relative clauses in the training corpus. When the network encounters the relative pronoun `who` it might predict an article as in (3) or a transitive verb as in

(4) `The boy who chased the dog hit the man.`

Once it encounters the noun phrase `the dog` in (3) it will predict the relative clause verb, yet another embedding, and so forth. Thus, the SRN predicts a probability distribution of all possible next elements and it lacks the *definiteness* of human sentence production. Unlike humans, the SRN is not generating a single target structure but conditional grammatical continuations. This non-determinism is not problematic when assessing SRN learning because model performance can be measured against the distributional properties of the training corpus. Ideally, the SRN has learned to perfectly reproduce the conditional probabilities in the input. In sentence production, however, performance should be measured against a specific target utterance, not against distributional properties of the learning environment. As the SRN approach to language processing suggests, how fast and how well a linguistic construction is learned may critically depend on frequency in the linguistic input (both at the level of individual structures and substructures). But in production, speakers go beyond reproducing input frequencies in that they select a definite structure (perhaps from several syntactic alternatives) to convey an intended meaning. In other words, a production event is not a function of the totality of linguistic experience but of ones current communicative goal, although the *ease* with which production is achieved may be a direct function of the former.[2] Hence, apart from the absence of semantic representations in SRN, the nature of

[2] It was argued in Ferreira (2003), however, that relative processing difficulty in humans does not solely depend on surface frequency.

SRN mappings indicates that they might not be well-suited for the study of production. These are fairly trivial observations about SRN and human behavior, which nonetheless might have important consequences for modelling language learning and processing.

It is a symptom of the lack of definiteness in SRN, for instance, that these models also learn sequences of words (or word categories) which are ungrammatical. In the above toy language example, the SRN learns that noun phrases are followed by transitive verbs or relative pronouns and that verbs are followed by noun phrases or superordinate clause verbs. Thus, in a trained state it might activate a sequence of word categories which corresponds to ungrammatical sentences such as

(5) `*The dog chases the dog hit the man.`

Since the SRN overlays statistical information from different input structures, it might develop representations that do not always allow the model to reject structures which are locally grammatical but ungrammatical globally.[3] While it is one of the main strengths of the SRN model to generalize by overlaying information from different input structures, the learned mappings might not be sufficiently constrained. Semantic representations which fix the models larger communicative goal (conveying a particular message) in a prediction sequence will constrain the learned mapping more tightly than local conditional expectations, because no meaning input in the model's experience corresponds to ungrammatical sentences such as (5).

In a neural network model definiteness requires that only the most active output units (words or categories) in a production sequence are evaluated in testing. An utterance is discounted as ungrammatical if the most active predictions mismatch the target word/category anywhere in the sequence. In order to achieve definiteness the model must in some way or other represent communicative goals, for example the intended meaning of an utterance. Such representations will change the learning task for the model, both algorithmically and conceptually. An SRN maps word sequences onto distributions of word categories, whereas a production model which represents meaning will map from messages to word sequences. The main task of an SRN is to extract the statistical regularities in the input to be able to predict items in a sequence. A production model with additional meaning input learns to transduce between data types, semantic representations and sentence forms. The statistical regularities of the training corpus may influence the model in learning a meaning-to-form mapping but to extract these regularities is not the primary task of such a model. Rather, the task of a production model is to find suitable mappings between messages and definite forms. This might force the model to develop representations which are quite different from SRN representations because in order to produce definite forms, input messages need to be separated more distinctly in hidden space than input structures in SRN learning. How well a meaning-form mapping for some structure is learned in such a model might be

[3] Whether this indeed happens will depend on the particular input language and training conditions. An example of such non-discriminatory behavior will be given in Chapter 7, drawn from a simulation by Lewis and Elman (2001).

modulated by distributional properties of the input, but it might also depend on other factors. Two structures could be semantically similar but very different in form and therefore hard to distinguish despite being frequent in the input. On the other hand, two structures could have overlapping semantic features and share subsequences of words in their sentence form and this might facilitate learning both structures even if they are infrequent in the input.[4] Language processing as meaning-to-form transduction, and the resulting definiteness of expression, are two important properties of the Dual-path model which the SRN lacks.

3.1.3 Symbolic generalization

A third limitation of the SRN approach lies in its exclusive reliance on experience to produce sentences. There is ample evidence that language learners are sensitive to the statistical regularities of the ambient language they hear at different levels of linguistic behavior. For instance, in speech stream segmentation children draw on transitional probabilities to learn word boundaries (Saffran et al., 1996; Johnson and Jusczyk, 2001). Distributional regularities in the linguistic environment support the acquisition of syntactic categories (Redington et al., 1998; Mintz, 2003) and frequency information can aid syntactic disambiguation (MacDonald et al., 1994). Statistical learning may therefore provide children with powerful mechanisms to structure linguistic experience and exploit the rich sources of information in the ambient language, but it may nonetheless be insufficient for language acquisition for two reasons. The non-linguistic environment is perceptually open, i.e., there is an unbounded variety of unprecedented event tokens that a speaker must be able to describe using familiar words in novel combinations. On the other hand, natural language itself is lexically open in that it can recruit novel words to fill the slots of familiar constructions. In addition, languages are combinatorial in structure, so no amount of experience covers their expressivity. Both forms of openness—linguistic and non-linguistic—require the generalization of linguistic knowledge beyond immediate experience and learning context and models of language acquisition should exhibit the flexibility to accommodate such generalization.

SRN with localist word representations are lexically specific (and so is the Dual-path model), and they cannot easily handle lexical openness. Moreover, it has been suggested that the SRN cannot use a familiar lexical item w in a learned syntactic frame $\mathcal{F}$ unless it has experienced the word w in $\mathcal{F}$ during training (Marcus, 1998). Marcus argues, for instance, that SRNs which are exposed to identity relations such as “a rose is a rose” or “a tulip is a tulip”, cannot generalize a novel word “blicket” to form the sentence “a blicket is a blicket” without exposure to this particular sentence. In the words of Marcus, SRNs cannot extrapolate beyond their training set, whereas humans have no difficulty in producing instances of the identity relation over novel words. If language processing employs symbolic capabilities, this form of generalization is not problematic.

[4] Effects of interference and facilitation between structures are discussed in more detail in Chapters 5 and 8.

Symbolic processing comprises at least two capacities, (i) the ability to bind instances to variables, and (ii) the use of variables in rules or primitive operations. A model of language processing which represents the identity relation as a template "a X is a X" and binds lexical items to these variable slots would be able to produce such structures over novel words. SRNs, on the other hand, do not develop 'abstract variable-based frames' (Chang, 2002) and cannot bind lexical instances to variable slots. The representations the SRN develops in the identity task are entirely shaped by experience and this causes difficulty in using them to generate novel sentences.

It was a major motivation for the Dual-path model architecture to augment a connectionist system (such as an SRN) with a variable-binding mechanism to investigate whether this mechanism would yield human-like symbolic behavior in the resulting hybrid system. At the same time it was aimed at retaining the statistical nature of connectionist learning in this system to balance symbolic processing with knowledge of the distributional regularities in the linguistic environment (cf. Chang, 2002, p. 610).

3.2 Features of the Dual-path model

The Dual-path model has three important architectural features which I will briefly describe in this section. The complete model configuration will be illustrated and explained in detail in the following section.

3.2.1 Separate pathways

The Dual-path model is a model of sentence production and syntactic development. As previously mentioned, it consists of an SRN extended with a system to represent sentence meaning. Both components—the SRN and the meaning system—are arranged to form separate information channels or *pathways*. One pathway, the standard SRN, learns distributional regularities over sequences of words and develops distributed representations of word categories. In sentence production, this *sequencing network* enforces constraints on the order of classes of sentence constituents.

The second pathway, called the *message-lexical system*, is a feedforward network which has hidden layers to represent the meaning of words and the thematic roles of event participants (e.g., AGENT, PATIENT, THEME, EXPERIENCER, RECIPIENT, GOAL, LOCATION, etc.). The message-lexical system learns to map the conceptual content of the sentence meaning onto corresponding word forms. It also learns to activate thematic roles in the right order within a sentence, a mechanism which will be explained below. Thus, at each position in a sentence, the message-lexical system activates sentence-specific, 'meaning-related possibilities' and the sequencing system activates 'syntactically appropriate possibilities' (cf. Chang, 2002, p. 622).

The two pathways of the Dual-path model intersect at the HIDDEN-layer of the sequencing network and at the output WORD-layer where they compete (or cooperate) to produce the next word in a sequence.[5] For example, early in learning the message-lexical system might activate verbs from different clauses in a sentence, e.g., a transitive and an intransitive verb, whereas the sequencing system activates all verbs in the transitive verb class. Since activation of both subsystems is summed at the output, the joint activation of a specific transitive verb will win this slot. Or, the message-lexical system might activate a transitive verb whereas the sequencing system activates a function word, e.g., a relative pronoun. In this case both pathways compete and the more active unit will win the slot. The subsystems learn different types of information, the integration of which ensures that words which are adequate to express the sentence meaning are produced at the right position of an utterance. The separation of these pathways is motivated by a study of sentence production (Bock and Cutting, 1992), which showed that sentence structure in production is influenced by distinct factors from lexical-semantic and syntactic processing.

3.2.2 The WHAT-WHERE division

In spatial processing of visual information, there are two tasks to be performed. Objects in the visual array need to be located and identified. Both tasks are dissociated. We can identify and categorize objects we perceive without having to describe their exact spatial position, and we can locate objects, e.g., obstacles in spatial navigation, without having to identify them first. Landau and Jackendoff (1993) suggested that this separation correlates with two different modules in the functional organization of the visual system in the brain, a 'what'-system for object categorization and a 'where'-system for object location. Both systems perform distinct subtasks but can temporarily be associated to locate a specific object, or to identify an object in a specific position. A system which could not perform these tasks independently would not recognize familiar objects in novel locations and would not be able to locate unfamiliar objects. But if both tasks, object identification and location, are functionally separate and can be linked together, the compound system can generalize to novel perceptual episodes.

In language production, familiar words can be used in novel sentence positions without a fundamental change in their lexical meaning. On the other hand, different words with distinct lexical meaning can assume the same position in a sentence. Thematic roles indicate the semantic relationship between the predicate and an argument of a sentence. Where in a sentence a word can be placed is constrained by which thematic role it can occupy, although the nature of such constraints is an issue of much controversy (see Goldberg, 2006). For instance, inanimate objects cannot usually occupy the recipient role in dative constructions, such as the ditransitive *He gave the dog the toy*, but they can always occupy the theme role (here: *toy*). The stipulated similarity between words in language processing and objects in spatial processing is that lexical meaning corresponds to object categorization and thematic role assignment corresponds to object

[5] I will use small capitals to denote layers of the Dual-path model throughout, and normal fonts to talk about neural network layers in general.

location. The thematic role of a constituent determines in which sentence position the constituent is placed (*where*) and the lexical meaning of a constituent determines who or *what* is placed there.

This analogy motivated the second important architectural feature of the Dual-path model. Different aspects of sentence meaning, such as concepts and thematic roles, are represented at different layers of the model. Concepts are locally represented by units in the so-called WHAT-layer and thematic roles are locally represented by units in another, physically and functionally distinct layer, the so-called WHERE-layer. As in spatial processing, concepts and thematic roles in these layers can temporarily be bound together by dynamically changing connection weights. Thus, it is possible to encode sentence meaning in such a way that different concepts (*what* information) can occupy the same thematic role (*where* information) and that the same concept could be assigned different thematic roles. Building the model, it was hypothesized that in this manner the Dual-path model architecture would inherit the generalization capabilities observed in visual processing.

3.2.3 Event semantics

Aspects of the Dual-path model are inspired by a construction grammar approach to language. By Goldbergs definition, constructions are form-meaning pairs which are neither compositional nor derivational (Goldberg, 1995). They can be atomic or complex, concrete or schematic, they have a holistic meaning and are basic linguistic units. For example, the mapping of the phonetic form [kæt] to the concept [CAT] is an atomic, concrete construction, and the mapping of the form [Subj V Obj Obl] to the semantic structure [X CAUSES Y TO MOVE Z] is a complex, schematic construction. The latter is the CAUSED-MOTION 'argument structure'-construction, which is instantiated by the sentence

(6) `He kicked the ball across the field.` (CAUSED-MOTION)

Many different verbs and prepositional phrases can feature in the CAUSED-MOTION construction, but it is not the syntactic properties of these constituents as listed in the lexicon which determine whether they are admissible arguments. Rather, constructions themselves select arguments which conform with their semantic and syntactic properties. Thus, non-prototypical verbs can be inserted into schematic constructions to yield innovative sentences such as

(7) `He sneezed the napkin off the table.`

The meaning of sentence (7) is not a function of its constituents (cf. Goldberg, 1995). The intransitive verb `sneeze` is used in an independent syntactic frame with its own meaning in which it becomes interpreted transitively. Argument structure constructions are schematic abstractions over event types which are basic to human experience, for instance CAUSED-MOTION, TRANSFER, or ACTING ON.

In the Dual-path model, sentence meaning is represented in the message-lexical system. A representational distinction is made between atomic concepts and complex constructions. Whereas the meaning of atomic concepts is locally represented by units in the WHAT-layer (and learnable connection weights to word forms), the meaning of complex constructions is represented by activation patterns in a dedicated EVENT SEMANTICS-layer. The event semantics marks the differences and, crucially, also the similarities between distinct constructions. For instance, there is a metaphorical similarity between the CAUSED-MOTION construction in its central sense

(8) X CAUSES Y TO MOVE Z: `Joe drove the car into the lake.`

and the TRANSFER construction

(9) X CAUSES Z TO RECEIVE Y: `Joe gave the present to Paula.`

Both constructions share the CAUSATION and MOTION event features. This similarity is reflected in the event semantics of the Dual-path model in that the same event roles are used to encode the meaning of both constructions. In particular, the goal (`the lake`) of CAUSED-MOTION in sentence (8) and the recipient (`Paula`) of TRANSFER in sentence (9) are collapsed onto the Z role in the event semantics. Representing similarities among constructions by shared features and similar patterns of activation in the event semantics, the model is able to generalize acquired knowledge from one construction to related constructions that share common event features. Examples of the event semantics message component in the model will be given below. The sensitivity of children and adults to event features of particular syntactic forms is well attested in the psycholinguistic literature (Fisher et al., 1991; Gropen et al., 1991; Kaschak and Glenberg, 2000).

3.3 Dual-path model architecture explained

In this section, the full Dual-path model architecture will be presented, and the two subsystems described in more detail. As was mentioned before, the sequencing pathway is a standard simple recurrent network as explained in the previous chapter.

3.3.1 Sequencing system

The Dual-path model maps meaning representations, which it receives as input, onto sequences of words. It learns in an error-based word-to-word prediction paradigm. The CWORD- and WORD-layers are the word input and output layers, respectively, of the model (Figure 3.2). In all experiments that will be discussed subsequently, an artificial English-like language of varying complexity was used. Lexical items in these languages, as well as morphological markers for tense, aspect and number, are represented locally at these layers. That is, each lexical item occupies exactly one node which signals the presence or absence of a word in a sequence by being switched on or off. Activation,

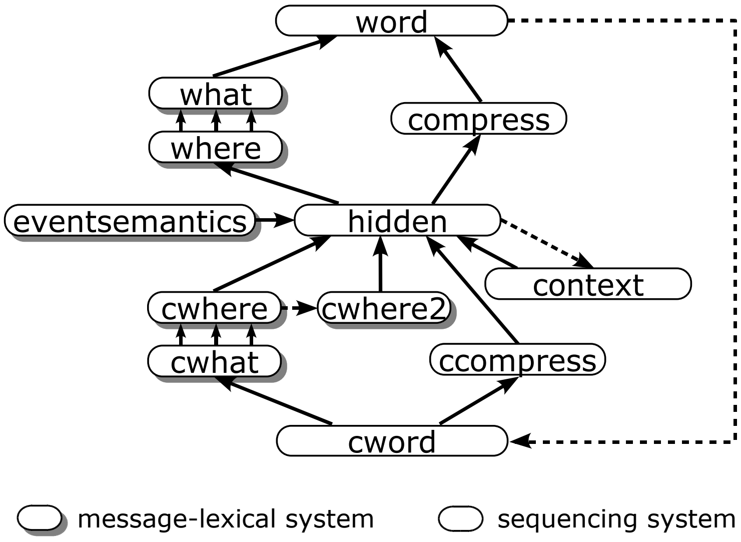


Figure 3.2: The Dual-path model architecture.

however, could be graded at the CWORD input, depending on the processing mode, and at the WORD output, due to the specific activation function used. Localist representations of one word by one unit may seem simplistic but they have the advantage that all lexical items viewed as vectors in input space are orthogonal by design. Thus, no semantic bias is induced by the experimenters' choice of features as in distributed lexical representations (cf. Page, 2000).

In the sequencing system, solid arrows in the diagram represent full learnable connectivity between layers and indicate the direction of activation spread. In other words, in a layer which is the source of a solid arrow, every unit is connected to every unit in the target layer, and activation can only propagate in the direction of the arrow. The strength of these connections is gradually adjusted in learning by backpropagation of error. Dashed arrows indicate copy-back connections along which the activation state of one layer is copied onto another. For instance, at each word in a sentence, the activation state of the HIDDEN-layer is copied to the CONTEXT-layer and fed back to the HIDDEN-layer at the next word. In this way, the CONTEXT-layer provides the HIDDEN-layer with a working memory system which enables the sequencing system to learn temporal contingencies in the input. Similarly, the dashed arrow between the WORD- and the CWORD-layer indicates that the model's word output in each sentence position is copied back to the input CWORD-layer on the subsequent time step. Thus, the model constantly monitors its own production output and predicts the next word based on the previous word plus its knowledge of the context of uttering this word, active in the HIDDEN-layer.

Between the lexical layers and the HIDDEN-layer, the sequencing system is equipped with special COMPRESS- and CCOMPRESS-layers (see Elman, 1991) which are roughly 1/3 the size of the HIDDEN-layer. These layers ensure that the model develops abstract representations of word categories instead of word-specific representations. This will be shown in more detail in Chapter 5. In addition to the activation from the CCOMPRESS-layer, the HIDDEN units receive input from the CONTEXT-layer which holds a copy of the previous time-step activation state of the HIDDEN units. Through the COMPRESS-layer bottleneck, the HIDDEN-layer then maps to the lexical output layer (WORD).

3.3.2 Message-lexical system

The message-lexical system is a feedforward network which holds the representations of sentence meaning in the Dual-path production model (Figure 3.2). This message is mapped to the same lexical layer to which the sequencing system projects.

To represent a sentence message, several special layers are sandwiched between the HIDDEN- and the WORD-layer in the message-lexical system, a semantic WHAT-layer and a thematic WHERE-layer. The WHAT-layer contains units which stand for concepts that represent the meaning of words in the lexical layers. This encoding is again localist, one unit represents one concept. In general, the WHAT-layer is smaller than the lexical layers because it only represents the semantics of all content words in the lexicon of the artificial language. For example, the WHAT-layer has units representing the meaning of verb stems and nouns, such as a unit [CHASE] and a unit [CAT], but no units representing inflectional morphemes, auxiliaries, prepositions or relative pronouns. The WHERE-layer contains units which stand for thematic roles that represent the semantic relationship between the verb and its arguments in a sentence. For example, the WHERE-layer has units representing the thematic roles of AGENT, PATIENT, RECIPIENT, as well as ACTION roles. Sentence-specific semantic content is represented through temporary *binding-by-weight* of the WHAT- and WHERE-layers. Such bindings are implemented by setting a connection weight between appropriate units in these two layers to a constant positive value. In this fashion, the WHAT units can represent the meaning of a word irrespective of the event role that word occupies in a particular sentence. Consider, for instance, the sentence `the cat chases the dog` in which `the cat` is the agent of transitive action. To represent the meaning of this sentence, the AGENT unit in the WHERE-layer would be bound to the [CAT] concept unit in the WHAT-layer by a connection weight. Similarly, the PATIENT unit in the WHERE-layer would be linked to the [DOG] concept unit in the WHAT-layer and the ACTION unit would be linked to the concept [CHASE]. Activation can then spread from the AGENT unit in the WHERE-layer to the lexical semantics [CAT] in the WHAT-layer, but not from the AGENT unit to the [DOG] unit, since these two units are not linked in the WHAT-WHERE-system for this particular sentence.

If `the cat` assumes a different thematic role in another sentence, e.g., the PATIENT role in `the dog chases the cat`, the [CAT] concept in the WHAT-layer would be dynamically bound to the PATIENT unit in the WHERE-layer to represent this aspect of sentence meaning. Hence, through the binding of roles and concepts, the message representation of the Dual-path model allows cats to fill different event roles in different sentences. At the same time, the role-independent representation of the lexical semantics of words in the WHAT-layer retains the common meaning that all occurrences of `cat` in different sentences share (cf. Chang, 2002, p. 619).

The second important aspect of the message-lexical system concerns the mapping from concepts to words. The WHAT- and WORD-layers are fully connected and the model has to learn a 'word tag' in the WORD-layer for each concept in the WHAT-layer. Once this mapping is learned, the model can then productively use this word in different thematic roles to describe novel events. On the other hand, the model can learn to use novel words in familiar syntactic frames because the intended mapping from concepts to word forms is independent of the specific event role 'location' which words occupy in a sentence. Thus, the acquired one-to-one mapping of meanings to words combined with the dynamic binding of concepts to roles in the WHAT-WHERE system enables the model to achieve lexical generalization (see Chapter 6).

To summarize the description of this part of the message-lexical system, Figure 3.3 depicts how the conceptual content of the intransitive sentence `the cat sleeps` would be represented in the WHAT-WHERE-system. The ACTION role 0A in the WHERE-layer is temporarily bound to the concept [SLEEP] in the WHAT-layer by a connection weight and the PATIENT role 0Y is temporarily bound to the concept [CAT] (the motivation for specific event role assignments will be given in paragraph 3.5.3 below). During training, the model learns to map the concepts [CAT] and [SLEEP] in the semantic WHAT-layer to the corresponding word forms `cat` and `sleep` in the WORD output layer.

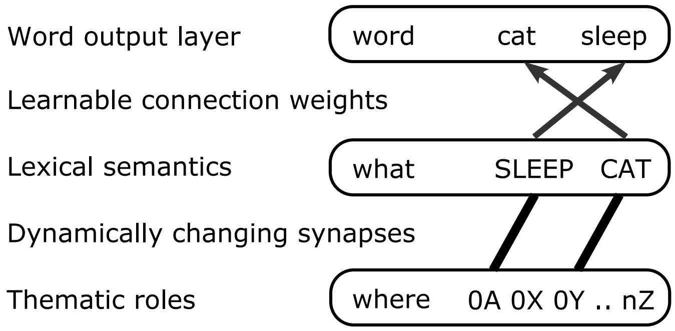


Figure 3.3: Message encoding of the intransitive sentence `the cat sleeps`.

The message-lexical system also contains an inverted copy of the WHAT-WHERE-system (Figure 3.2, page 56). This subnetwork consists of the CWHAT- and CWHERE-layers, the analogues to the WHAT- and WHERE-layers. These layers are sandwiched between the CWORD- and HIDDEN-layers in reverse order and can be thought of as a comprehension counterpart to the WHAT-WHERE system. When the model produces a lexical item at the WORD-layer, this word is fed back to the model's CWORD-layer on the next time-step. This CWORD input is mapped to concepts in the CWHAT-layer in a comprehension direction. The model has to learn this mapping in order to make sense of its own production output. The CWHAT-layer again has dynamic bindings with the thematic roles in the CWHERE-layer. These bindings are preset in identical fashion to the WHAT-WHERE bindings before sentence production begins. The purpose of this subsystem is to inform the HIDDEN-layer at each point in processing about which thematic role a previously produced word occupied. Consider the sentences `the cat chased the dog` and `the cat sleeps` and suppose the model has just produced the constituent `cat` at the WORD-layer. Then this word is copied back as input to the model at the CWORD-layer. For the network it could be the beginning of the transitive or the intransitive structure. In order to produce the intended structure the HIDDEN-layer needs to activate the correct sequence of roles in the WHERE-layer. Hence, the model needs to know the event role of `cat` in the conceptual structure of the target sentence, otherwise it does not know how to continue from here. The reversed WHAT-WHERE-system which feeds into the HIDDEN-layer delivers precisely this kind of information. Once `cat` is fed back to the CWORD-layer this will activate the concept [CAT] in the CWHAT-layer (provided the model has already learned this comprehension mapping). Since [CAT] is dynamically linked to a specific thematic role in the CWHERE-layer, activation spreads along this connection and switches on a role unit which informs the HIDDEN-layer about the thematic role of the previously produced word (`cat`). In addition to the CWHERE-layer, there is a CWHERE2-layer which functions similar to the CONTEXT-layer working memory in the sequencing system. At each point in time the CWHERE2 units sum the current activation state of the CWHERE units and the previous activation state of the CWHERE2 units. In contrast to the CWHERE units, which activate only the most recent role, the CHWERE2 units provide the HIDDEN-layer with a time-averaged history of previously activated roles.

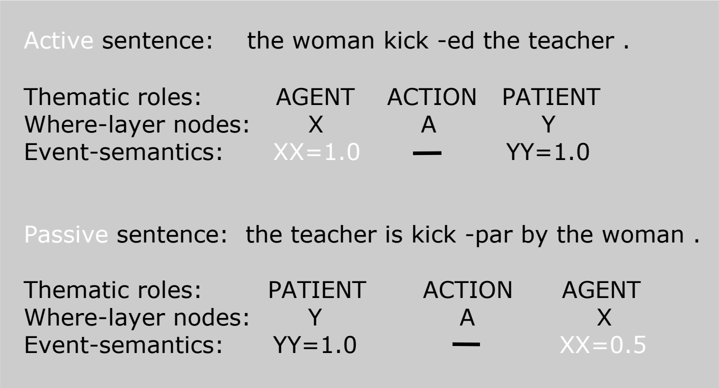
The CWHERE- and CWHERE2-layers can only accurately inform the HIDDEN-layer about previously produced roles if the model has learned the correct CWORD-CWHAT mapping, i.e., if it understands the meaning of the words it uttered. The error signal which is backpropagated from the WORD-layer downwards, however, rapidly decays as the number of layers increases over which the error is distributed. It is too weak to adjust the CWORD-CWHAT connections during learning. Hence the CWHAT-layer received its own error signal from the previous time-step WHAT-layer activation state. Initially, the target activation state of the WHAT-layer is not very distinctive because the model does not activate the correct WHERE roles at the beginning of training. It becomes more distinct and reliable over time only if the model learns to sequence WHERE roles appropriately. Thus, the assembly of inverted layer duals in the message-lexical pathway creates a learning problem with non-trivial dependencies. Consider again the sentence `the cat chased the dog`. In order to produce the intended active transitive construction, the model must perform two main tasks in the meaning system, (i) sequence appropriate thematic roles in the right order at the WHERE-layer, and (ii) map the conceptual content to correct word forms. That is, for example, towards the end of the sentence the model must activate `the dog` as the patient of the event in the message and map [DOG] to `dog`. To learn (ii), the lexical semantics of `dog`, the model must learn (i), role sequencing, because role sequencing precedes the meaning-to-form mapping in the production process. Task (i) is controlled by the HIDDEN-layer whose activation state relies on information coming from the CWHERE-CWHERE2 system about previously produced roles. The signal from this subsystem derives from the CWHAT activation state which is a function of how well the model comprehends the words it hears from itself. As outlined above, however, comprehension at the CWHAT-layer depends on the activation state of the WHAT-layer which in turn depends on role sequencing at the WHERE-layer. Hence, role assignment and word production/comprehension are tightly coupled. The model learns to identify the thematic roles that words occupy in the semantics of a sentence it generates itself, and concurrently it learns the meaning of words. In this way, the model "bootstrap[s] word learning, by incrementally learning to comprehend the previously produced semantics" (Chang, 2002, p. 620). There is evidence that adults assign thematic roles incrementally in sentence comprehension (Sedivy et al., 1999) and that observed events actively influence this incremental assignment (Knoeferle et al., 2005). In learning the meaning of words, children must infer intended referents of words in the speech they hear and assign thematic roles in observed events. To draw on visual information in an environment shared with the speaker, they require selective attention mechanisms which guide their attention to relevant aspects of the visual field. They also require joint attention capabilities to pick out intended referents in word learning. Evidence from the work of Tomasello (1999, 2003) indicates that children have both these required abilities. This suggests that the concurrency of incremental role assignment and word learning in the Dual-path model does not rest on assumptions which might be beyond the capacities of language learning children.

3.3.3 Event semantics-layer

The complete message-lexical system of the Dual-path model comprises the two WHAT-WHERE-systems plus an EVENT SEMANTICS-layer which projects directly into the HIDDEN-layer (Figure 3.2, page 56). The event semantics holds information about the type of event described by the target utterance. Specifically, it provides information about the number of participants in an event. For this purpose the EVENT SEMANTICS-layer consists of units, or event features, which signal the presence of participants. An active XX feature, for instance, would signal the presence of a causer or an agent, a ZZ feature would signal the presence of a goal, recipient or location. Depending on the artificial language and learning conditions, the event semantics can have additional features, e.g., a feature AA for the action in an event, a feature DD for prepositions, features PAST, PRES for tense, and SIMP, PROG for aspect. For a multi-clause language with a complex syntax, additional features are required to signal the relative prominence of basic events, or to specially mark other participant features. The following chapter is devoted entirely to optimizing the representations in the event semantics of such a language for learning and generalization. To give an example, if the event is an instance of the AGENT-PATIENT construction, typically expressed by an active transitive sentence such as `the woman kicked the teacher`, the event semantics would signal the presence of an agent by activating the feature XX, the presence of a patient by activating the feature YY, and (optionally) activate the action feature AA, the past tense feature PAST, and the simple aspect feature SIMP. This pattern of activation would be fully set in the EVENT SEMAN-

TICS-layer prior to sentence production and remain active without change in the course of production.

The event semantics provides the conceptual structure of an event which the model experiences while trying to predict the next word in a sequence. It constrains the model's output in that it encodes the number of participants in the target utterance and their semantic relationship. Moreover, the event semantics specifies the relative prominence of participants in an event. Higher activation of one feature over another signals the higher prominence of the corresponding participant. The more prominent a role feature is, the sooner the model produces the word associated with this role in the output sequence. Consider the sentence from above in passive voice `the teacher is kicked by the woman`. The two sentences are not distinguished in terms of the number of participants, their semantic roles, tense or aspect, or their conceptual content in the WHAT-WHERE-system. To encode the distinction between active and passive structures in the event semantics, the prominence of the YY feature relative to the XX feature can vary. This is implemented through a reduction of activation of the XX feature relative to its default level (see Figure 3.4). In other words, if an active structure is intended to express the message, both agent and patient features in the event semantics have the same level of activation. If the passive construction is intended, the agent feature XX is reduced relative to the patient feature YY. This biases the model towards activating the Y role prior to the X role in the WHERE-layer and hence to produce `the teacher` before `the woman` in the word output sequence.



Active sentence: the woman kick -ed the teacher .

Thematic roles:	AGENT	ACTION	PATIENT
Where-layer nodes:	X	A	Y
Event-semantics:	XX=1.0	—	YY=1.0

Passive sentence: the teacher is kick -par by the woman .

Thematic roles:	PATIENT	ACTION	AGENT
Where-layer nodes:	Y	A	X
Event-semantics:	YY=1.0	—	XX=0.5

Figure 3.4: Marking the active/passive distinction in the event semantics.

In visual processing, aspects of the perceived scene are organized into figure and ground. Which aspect is categorized as figure and which as ground depends on the attentional focus. Reversing the attentional focus, one aspect of a scene can alternate between figure and ground. The way the active/passive alternation is encoded in the event semantics is analogous to the figure-ground alignment in visual perception. Differential activation of features in the EVENT SEMANTICS-layer encodes the relative prominence of corresponding thematic roles in the message. By reducing the XX feature activation in the passive construction, the model's attention is focused on the patient role Y, thus bringing it to the foreground and forcing the X role, corresponding to the reduced feature XX, into the background. This principle is also used to encode other syntactic alternations in the language. It is important to point out, however, that the event semantics does not provide the model with syntactic frame information because the features in the event semantics do not map onto syntactic roles one-to-one. The model develops its syntactic representations through learning.

The Dual-path model is endowed with semantic role variables in the WHERE-layer. These variables become instantiated through fast-changing weight bindings with concept nodes in the WHAT-layer. This mechanism was designed to enable generalization of novel constituents to familiar syntactic frames. It is not sufficient for symbolic behavior, however, to equip a connectionist system with such variables unless the system is able to properly use these variables. The Dual-path model can use these variables if it learns to activate the appropriate WHERE-layer roles of the sentence message in the right order, and has learned the lexical semantics for the WHAT-layer concepts temporarily bound to these roles. The EVENT SEMANTICS-layer feeds directly into the HIDDEN-layer of the sequencing system and guides the use of the role variables. It provides information about which structure the sequencing system should select in order to produce all variables in the message. Activation differences between features in the EVENT SEMANTICS-layer help the model activate the correct sequence of role variables. In the words of Chang, Dell, and Bock (2006), "the event semantics helps the sequencing system learn language-specific frames for conveying particular sets of roles by giving the sequencing system information about the number of arguments and their relative prominence" (p. 242). Without event semantics the Dual-path model does not have access to the intended message. Consequently it cannot exploit sentence-specific semantic features to constrain the sequencing process. Its word predictions are based on previously produced words only. The Dual-path model without event semantics still has thematic role variables in the message-lexical system and acquires syntactic frames in the sequencing system, but was nonetheless shown inferior to the Dual-path model with event semantics in terms of a variety of generalization tasks (cf. Chang, 2002, pp. 625). This suggests that variables alone are not sufficient architecturally to exhibit symbolic generalization unless the use of these variables is supported appropriately.

3.4 Production sample

The Dual-path model architecture is quite complex, and so is the flow of information during learning and production. It will therefore be helpful to walk through the production process step-by-step by means of a concrete example. Consider the prepositional dative sentence `the girl give -s a toy to the cat`.[6] It is assumed that the model is trained and produces this sample sentence correctly. I will focus on the quintessential processes in a model subject which has learned to produce sentences in the intended way. Thus, I will describe an idealized production event.

[6] Note that the present tense morpheme `-s` is treated as a separate lexical item.

Before production begins the model is initialized. All learnable connection weights in the network are set to those values that developed in the model during training. The training procedure is described in more detail in the appendix and the subsequent chapters. The CONTEXT-layer units are initialized to a value of 0.5. Next, the complete message for the target sentence is set in the message-lexical system prior to production and this message will not be manipulated externally in the production process. The message associated with the target sentence is shown in Table 3.1. First, all thematic roles in the message are linked with the appropriate concepts by creating synaptic connections between the corresponding units in the WHERE- and WHAT-layers. These connections are set to a fixed arbitrary value which is sufficient to ensure that an active WHERE-layer node activates the target WHAT-layer node. Furthermore, the value of these connections needs to be such that both pathways in the model can make a balanced contribution to the overall production process. A smaller value will give more weight to the sequencing system, a larger value more to the message-lexical system. In the above example, the agent unit X in the WHERE-layer gets linked with the concepts [THE] and [GIRL] in the WHAT-layer, the action node A is linked with the verb stem concept [GIVE], the theme unit Y is linked with the indefinite article [A] and [TOY], and the recipient unit Z is linked with [THE] and [CAT]. The same sentence-specific connections are set in reverse direction between the CWHAT- and the CWHERE-layers of the model, the CWHERE2-layer is initialized like the CONTEXT-layer. Secondly, the event semantics component of the message is set. In the example, the XX and YY features are switched to 1.0, and the ZZ feature is switched to 0.5. This pattern of activation signals to the sequencing system that a TRANSFER construction with three participants is intended and the activation value of the recipient feature, which is lower than baseline, biases the model towards expressing the message with a prepositional dative instead of a ditransitive sentence. If the ditransitive alternation `the girl give -s the cat a toy` had been intended the recipient feature would have been set to 1.0. As in case of the active/passive alternation differential activation encodes the relative prominence of roles, and a feature reduction takes the corresponding role out of the attentional focus. Finally, the present tense feature PRES and the simple aspect feature SIMP are turned on in the EVENT SEMANTICS-layer.

Sentence	`the girl give -s a toy to the cat`			
Thematic roles	X (AGENT)	A (ACTION)	Y (THEME)	Z (RECIPIENT)
Concepts	THE, GIRL	GIVE	A, TOY	THE, CAT
Event semantics	XX=1.0	PRES, SIMP=1.0	YY=1.0	ZZ=0.5

Table 3.1: Message components for the target utterance.

The Dual-path model operates on a discrete time scale, the production event is subdivided into *ticks* and each tick corresponds to one word. After the model is initialized and the message is set in the described way production starts on the first tick with no word input. Activation spreads from the uniformly active CONTEXT-layer and from the

EVENT SEMANTICS-layer to the HIDDEN-layer of the sequencing system. Connections between these layers have been trained, hence the HIDDEN-layer activates the agent role X in the WHERE-layer and consequently both concepts [THE] and [GIRL] will be activated in the WHAT-layer. Activation also spreads from the HIDDEN-layer to the COMPRESS-layer where the sequencing system represents grammatical word categories (Chapter 5). Since the model has learned the meaning of words, the message-lexical system activates the lexical items `the` and `girl` at the WORD-layer. Both lexical items become available for production. The COMPRESS-layer, however, activates only the determiner category because all sentences in the input started with either a definite or indefinite article. The activation from both pathways is summed, thus the sequencing system enforces the article and the model produces `the` on the first tick.

After the definite article is produced, `the` is copied back to the CWORD-layer on the second tick. Now activation spreads through both pathways to the HIDDEN-layer. The previous time-step activation pattern of the HIDDEN-layer is supplied by the CONTEXT-layer as an additional input. In the comprehension system of the model, `the` is mapped to the concept [THE] which activates the X role in the CWHERE-layer because the produced article is part of the noun phrase that is associated with this role. Similarly, the Z role in the CWHERE-layer is active because the noun phrase associated with Z also uses the definite article. The CWHERE2-layer, on the other hand, is silent because the CWHERE-layer was inactive on the first tick. In the sequencing system, the article is mapped to the determiner category in the CCOMPRESS-layer. In the input language, articles were always followed by nouns, hence the HIDDEN-layer activates the noun category in the COMPRESS-layer. From the message-lexical system, the HIDDEN-layer receives ambiguous information since both the agent role X and the recipient role Z are activated by the definite article at the CWHERE-layer. Consequently, the sequencing system is uncertain which role should be activated next in the WHERE-layer. At this point in processing, statistical information from the training corpus helps the model in selecting a role. Passive datives, such as `the cat is given a toy by the girl`, were absent from the learning environment in the condition from which this processing example is drawn. Because the dative structure itself is signaled by the event semantics, and the model never experienced a dative with a sentence-initial recipient it opts against the Z role and strongly activates the agent role X at the WHERE-layer instead. Again, the concepts [THE] and [GIRL] are activated at the WHAT-layer and therefore the corresponding lexical items at the WORD-layer. This time, however, the COMPRESS-layer does not support articles but favors all nouns at the WORD output. Therefore the model produces `girl` on the second tick.

On the third tick, `girl` is copied to the CWORD-layer and the model continues the production process. In a trained state, the model comprehends its own word predictions. Thus it activates [GIRL] at the CWHAT-layer which now unambiguously identifies the previous role as the agent role X in the CWHERE-layer, due to the preset synaptic bindings between these layers. This causes the HIDDEN-layer to sequence the action unit A next at the WHERE-layer and triggers the activation of the concept [GIVE] at the WHAT- and ultimately the transfer verb `give` at the WORD-layer. This prediction is in accord with the word class prediction in the sequencing system, so give gets produced. A change to this routine occurs on the fourth tick. The CWHERE-layer signals to the HIDDEN-layer that the action role A has been sequenced on the previous tick. According to the event semantics the intended message conveys a non-continuous event in the present. The knowledge to activate the tense marker -s resides in the sequencing system because the action role A is linked only to the verb stem in the WHAT-WHERE-system. Thus, the message-lexical system stays largely inactive and the COMPRESS-layer activates the tense morpheme at the WORD output.

After the verb form is complete the model needs to decide whether to express the message with a prepositional or a double object dative. The choice is between sequencing the theme role Y or the recipient role Z next. The previously produced words and roles are neutral between these options. Moreover, in either case the sequencing pathway must activate the determiner category next. At this juncture between syntactic alternations, the HIDDEN-layer relies exclusively on the semantic information from the EVENT SEMANTICS-layer which biases the model towards the prepositional dative rendering of the message, because the YY feature is more active than the ZZ feature. Hence, the model produces the indeterminate article a on the fifth, and the noun toy on the sixth tick. The structural choice between syntactic alternations is made and the sequencing system generates the preposition to after the theme of the prepositional dative construction. Since function words such as prepositions are not encoded in the message, the sequencing system draws on learned statistical regularities to produce the preposition. It is aided by the CONTEXT-layer which supplies the activation state of the HIDDEN-layer on the previous tick and this 'sequential context' informs the sequencing system that the dative theme is complete and a prepositional phrase should be produced.

When to is fed back to the CWORD-layer, the CWHERE-layer stays silent, because the preposition is not thematically or conceptually represented in the WHAT-WHERE-system. The CWHERE2-layer, however, has recorded all roles which have been produced so far. This is due to the time-averaging of the previous activation state of both layers in the CWHERE2-layer. Thus, it signals to the HIDDEN-layer that the agent, action and the theme role have been sequenced already. The model uses this cumulative history alongside the constructional information from the event semantics to activate the recipient role Z at the WHERE-layer on the next two ticks. This leads to the production of the cat. On the final time step all units representing thematic roles in the intended message are fully turned on at the CWHERE2-layer. The WHERE-layer roles are almost completely silent. The model knows that all participants which are encoded in the event semantics have been generated and the sequencing system produces the period symbol '.' to mark the end of the sentence.

The Dual-path model production process can be summarized as follows. Words are incrementally activated from the sentence-specific semantic content in the message-lexical system. The sequencing system constrains this process by enforcing the grammaticality word category sequences, and provides the functional constituents in the target utterance. This system is guided by statistical regularities in the learning environment and the conceptual structure of the intended message in the event semantics. At the output layer there is a competition between the message-lexical and the sequencing pathway for the subsequent word slot. The interplay of different kinds of representations in the separate pathways—the sentence-specific content of the message-lexical system and the abstract syntactic frames in the sequencing system—is the key to felicitous sentence production in the model.

The Dual-path model does not implement any particular psycholinguistic theory of sentence production. Yet, the model's production process can vaguely be compared with one of the most influential such theory by Levelt which pictures sentence production as a linear progression of successive stages: conceptualization, formulation, articulation, and self-monitoring (Levelt, 1989). In the Dual-path model setting the preverbal message in the meaning system can be viewed as the conceptualization stage. Conceptualization is achieved by a planning system which, strictly speaking, is external to the model. One aspect of formulation in Levelt's theory is grammatical encoding to select a sentence surface form. Roughly, this stage corresponds to the model utilizing the event semantics to chose between syntactic alternatives which express the intended message. In the Dual-path model, though, grammatical encoding is accomplished incrementally during production and does not precede articulation. The model only selects a syntactic form at structural choice points in a sentence, not prior to production. The incrementality of grammatical encoding in the model will be demonstrated in Chapter 5. Articulation occurs in the model when activation spreads along both pathways and a word is selected based on competition or cooperation between the two functionally independent subsystems. This output is fed back to the model in the self-monitoring stage in which it is processed in a comprehension direction. Lexical meaning and thematic roles are assigned to the monitored constituent to guide further production of the target word sequence.

3.5 Model assumptions

Experimental results obtained by computational modelling should always be evaluated against the strongest assumptions underlying the model. It is therefore crucial to put these assumptions in plain view, so that the significance of results can be assessed without in-depth study of the model itself. I follow the classification of assumptions from Table 1 in Chang et al. (2006, p. 240) and the discussion therein.

3.5.1 Learning assumptions

A fundamental assumption underlying the Dual-path model (and many other connectionist models) is that language learning occurs *qua* processing. There are three aspects to this *learning-as-processing* assumption.

Learning-as-processing First, it is assumed that language acquisition is an instance of implicit learning. Knowledge of a language develops incrementally by fine-tuning the processor to the task of mapping semantic representations onto appropriate sentence forms. This knowledge is non-declarative and not accessible to deliberate recall. In the Dual-path model, experience-driven implicit learning corresponds to adjusting the connection weights in response to the statistical contingencies in the training environment. Secondly, implicit knowledge develops gradually in the very processor itself. There is no special mechanism external to the processor which provides syntactic knowledge in sentence production. The production system itself embodies this knowledge and it is only manifest in performance. The Dual-path model does not induce syntactic knowledge from training patterns to store this knowledge in a dedicated memory. Rather, syntactic knowledge develops in the connectivity between layers in the processor. Even though the model is structured, all syntactic knowledge resides in a single set of connection weights which is the functional core of learning, knowledge representation and processing in the model. This conception of learning and processing has been advocated in several psychological models (e.g., Gupta and Cohen, 2002, Botvinick and Plaut, 2004). And third, the Dual-path model incorporates a continuity assumption. It is assumed that the mechanisms which drive language acquisition continue to function from childhood through adulthood. In the network, individual nodes saturate during learning and synaptic plasticity decreases. Thus, it becomes increasingly inflexible over time. Nonetheless, the model's basic learning and processing mechanism remains essentially the same at all stages of development.

Prediction error The Dual-path model learns in a situated comprehension mode, in which it receives external linguistic input and has access to the meaning of these utterances. At each word position, it predicts the next word in the overheard utterance. The difference between the internal predictions and the external input is then used to adjust the strength of the model's connection weights. Thus, it is assumed that human language learners engage in word prediction during comprehension and are sensitive to mismatches between internal predictions and external sentence input. Pickering and Garrod (2007) argue that prediction and imitation in the production system is used in comprehension and combined with the linguistic input in a dynamic way. On their view, the rapidity of human comprehension can be explained if the comprehension system recruits the production system to emulate external production through covert prediction. Evidence for prediction in comprehension comes from a number of psycholinguistic studies of sentence processing (Altmann and Kamide, 1999; Knoeferle et al., 2005) and evidence for the activation of the production system during comprehension has been found in neuro-imaging studies of speech perception (Watkins et al., 2003; Heim et al., 2003).

MacWhinney (2005) has argued that prediction error might be the basis of powerful learning mechanisms in language acquisition. In comprehension, the learner compares word predictions from her own production system with the word sequence she hears. Detected discrepancies between predictive expectations and actual input are utilized to learn syntactic structure. In production, the learner monitors her own word output to recover from structural errors and overgeneralization. MacWhinney suggests that these learning strategies—receptive and expressive monitoring, respectively—might jointly overcome the 'logical problem of language acquisition'.

3.5.2 Architectural assumptions

The architectural assumptions which the Dual-path model was built on have partially been motivated before. I briefly recall them here.

Two pathways The model has separate meaning and sequencing systems which compete for the next word slot at the output layer. A growing body of evidence suggests that there is a dissociation between linguistic capacities associated with lexical-semantic and syntactic knowledge. This evidence supports theories according to which distinct aspects of language processing are subserved by different neurocognitive systems, or even different cortical areas. Ullman (2001) provides a survey of this evidence for 'dual-system' models with respect to grammatical processing and lexical memory from several domains such as neurological and developmental disorders, functional imaging, and psycholinguistics. The Dual-path model is consistent with these findings in that knowledge of syntactic frames and lexical semantics are acquired in functionally and physically distinct subsystems.

SRN sequencing In the Dual-path model, sequencing is accomplished by a simple recurrent network (Elman, 1990, 1991) which learns temporal contingencies in the input by means of a simplified working-memory system (context). SRNs have been used successfully in many cognitive domains which require learning sequential structure such as artificial grammar learning (Cleeremans, 1993), discrimination learning (Christiansen and Curtin, 1999a), routine action performance (Botvinick and Plaut, 2004), and natural language processing (see Christiansen and Chater (1999a) for an overview). Thus, it has been demonstrated in many domains that SRNs exhibit a similar sensitivity to sequential dependencies as human subjects.

3.5.3 Representational assumptions

There are two important assumptions underlying the Dual-path model which govern the way linguistic information is represented in the model.

WHAT-WHERE **separation** The conceptual content of a sentence is represented by dynamic bindings between concepts in the WHAT-layer and thematic roles in the WHERE-layer. The notion of fast-changing synaptic binding through spontaneous synaptogenesis is not well supported neuroscientifically, but for my purposes it is not critical how theses bindings are realized. What is important are the computational properties of representations that connect two separate systems with temporary bindings, regardless of the specific biophysical realization of binding in the human brain. These representations are motivated by the functional distinction between object recognition and spatial location in visual processing described in Ungerleider and Mishkin (1982). Through lesioning of the brains of macaque monkeys, they discovered a bifurcation between two anatomical pathways in the primate visual system, the *ventral* and the *dorsal* stream. The former is involved in the identification of objects, while the latter communicates the location of objects. Generalizing functional mappings from monkeys to humans is not unproblematic. Neuroimaging experiments, however, also support the idea of functionally dissociable processing systems for object recognition and spatial location in in human visual perception (Haxby et al., 1991). Several other neuroimaging studies have since strengthened the case for functionally and anatomically distinct processes in the visual system that encode 'what' and 'where' information (Smith et al., 1995; Mecklinger and Pfeifer, 1996; Mecklinger and Müller, 1996).

XYZ roles A second representational assumption concerns the assignment of thematic roles to sentence constituents. The Dual-path model of Chang, Dell, and Bock (2006) uses three thematic roles—X, Y, and Z—to encode the semantic relationship between event participants. These roles are represented by units at the WHERE-layer and by role features in the EVENT SEMANTICS-layer. The X role is assigned to agents, causes and stimuli, the Y role to patients, themes and experiencers, and the Z role to goals, locations, recipients and benefactors. This encoding does not follow any specific linguistic theory of thematic roles, although it combines aspects from several such theories (Dowty, 1991; Goldberg, 1995; Levin and Rappaport Hovav, 1995). The prime motivation for the XYZ role encoding, according to Chang et al. (2006), was to develop a thematic role representation which reflects the sequential process of scene analysis and approximates the movement of attention in visual processing. Evidence from several studies of language processing suggests that both comprehension and production are influenced by selective attention to spatial regions in observed scenes (Griffin and Bock, 2000; Knoeferle et al., 2005).

In the XYZ role representation, the Y role is assigned to the event participant which is most saliently affected, moved or changed by the action in an event. In the artificial languages I will mainly be using, this includes the subject of intransitives (`the cat` in the sentence `the cat sleeps`) and transitive objects (`the dog` in `the cat chases the dog`). It also includes objects which are transferred in dative events (`the toy` in `the cat gives the toy to a dog`). The X role is assigned to transitive and dative subjects only. The Z role is assigned to recipients in datives and oblique objects (`the boy` in `the cat runs with the boy`). Perhaps the most unconventional feature of the XYZ representation is that intransitive agents and transitive patients are assigned the same Y role. This treatment was motivated in Chang et al. (2006) by a study of Goldin-Meadow and Mylander (1998) on gesture in deaf children of non-signing parents. It was found that these children tend to gesture about intransitive agents and transitive patients before gesturing about actions which Chang et al. (2006, p. 241) interpret as evidence that “there is a prelinguistic basis for treating them as the same role”. I adopted this role assignment convention throughout.

3.6 Past research with the Dual-path model

The Dual-path model was first introduced in Chang (2002). In this paper, it was investigated to what extent the model exhibits symbolic generalization capacities. For instance, it was tested whether the model was able to produce familiar words in novel slots, whether it could produce the identity frames mentioned in Section 3.1.3, and whether it could produce novel adjective-noun pairs. By comparing the Dual-path model to similar production model architectures, it was shown that the model was superior in terms of these generalization tasks. In this way, it could be demonstrated that several Dual-path model features were essential for generalization. The binding-by-weight feature in the WHAT-WHERE-system, for example, enabled the model to produce familiar words in novel slots, whereas a binding-by-space message which used similar thematic roles was not satisfactory. It was also shown that thematic role variables were not sufficient to produce identity frames (*a blicket is a blicket*) unless the model was equipped with an EVENT SEMANTICS-layer which guided the use of these variables. Moreover, it was shown that the separation of pathways was conducive in all generalization tasks. Linking the pathways contaminated the representations in the sequencing system with lexical-semantic information and prevented the Dual-path model from learning fully abstract syntactic frames. Since generalization is a critical aspect of language acquisition, and therefore a benchmark for models of syntactic development, this model comparison provided computational support for some of the basic architectural and representational assumptions behind the Dual-path model. The dual pathway assumption was independently supported by a demonstration of the model’s ability to account for some double dissociations in aphasics—e.g., between function and content words and light and heavy verbs—when lesioned selectively in either of the two pathways.

While the Chang (2002) study was mainly concerned with motivating the Dual-path model architecture and its representational assumptions in a wide range of generalization tasks, the study of Chang et al. (2006) focussed on testing whether the model was able to account for sentence processing data, both in children and adults. In this study the model was applied to explain experimental data from three methodological paradigms which tap into the use of syntactic representations: structural priming, elicited production, and preferential looking. Speakers are inclined to reuse syntactic structures across sentences they produce. For instance, when exposed to a prepositional dative sentence prior to describing a pictured event, people are more likely to use a prepositional dative in that description rather than the double-object dative which could be used to express the same meaning. This effect is called syntactic priming (Bock, 1986; Bock and Loebell, 1990). The Dual-path model approach hypothesized that syntactic priming is a form of implicit learning. Priming in the model resulted from small changes in the connection strength between units of the learning mechanism during the processing of the prime structure, and Chang et al. (2006) were able to qualitatively match and explain a variety of structural priming data in adults using the model. Priming in the model, for example, was persistent over lag (filler sentences), was insensitive to lexical and morphological overlap between prime and target, and differentially sensitive to meaning, depending on the prime-target alternation. These findings were consistent with human priming data from a number of previous studies (see Chang et al., 2006, p. 263 for details).

In the same study, the Dual-path model was tested on how well it matched data from language acquisition in several tasks. Chang, Dell & Bock found, for instance, that structural priming occurred also during syntactic development not merely in an adult state at the end of learning. Furthermore, preferential looking preceded elicited production in the model when tested on transitive frames. Transitives preceded intransitives in preferential looking and the transitive construction developed in a verb-specific way in production. Again, these results were broadly consistent with what is known on these issues in developmental psycholinguistics.

Most recently, the Dual-path model has been used in a cross-linguistic study of English and Japanese (Chang, 2008). Since the semantic features in the model's message are generic and the model learns to map messages onto grammatical forms, the model could be applied to languages other than English. In this study, it is demonstrated that the Dual-path model was able to learn artificial English- and Japanese-like languages and displayed language-specific behavior which matched human behavior with respect to heavy NP shift and lexical accessibility in production. For instance, the model exhibited the short-before-long NP preference in English and the long-before-short NP preference in Japanese, and thus could explain shift direction in both languages in a single mechanism. In addition, the model showed a preference to order animate before inanimate NPs in English and Japanese because it was sensitive to distributional properties in the input which support this ordering.

To summarize, the Dual-path model is tying together language learning and sentence production which have traditionally been studied separately in psycholinguistics. Since learning and processing take place in the same mechanism, the model offers a unified approach to acquisition and adult production. In the three studies of Chang (2002, 2008) and Chang et al. (2006), the Dual-path model was tested in a wide variety of acquisition, generalization and processing tasks. The findings indicate that the model's behavior is largely consistent with human performance in these tasks. This provides converging evidence that the model captures important aspects of human syntactic development and sentence production in a formal theory of learning and processing. In the remainder of this thesis, I will attempt to adduce further support for this claim from the domain of complex natural language syntax.

Chapter 4

Learning

In this chapter I describe the basic extension of the Dual-path model architecture necessary to accommodate the production of multi-clause utterances. Several meaning representations for complex sentences with relative clauses are presented. Learning results for these message types are discussed and analyzed. I identify an optimal set-up which will form the basis for simulations in subsequent chapters.

4.1 Introduction

The Dual-path model of Chang (2002) and Chang et al. (2006) has been used to investigate the processing of single-clause utterances exclusively. In language acquisition, however, many controversial claims about the limits of data-driven learning and the role of universal grammar are intimately tied to complex sentence structure, for example, the purported innateness of the human capacity for recursive productivity (Hauser et al., 2002) or the non-learnability of structure-dependent rules of grammar (Crain and Pietroski, 2001). In theories of language production and comprehension, complex sentence structure plays a critical role in elucidating the nature of syntactic processing. Often, fundamental properties of the human language system are inferred from the differential processing of structurally distinct multi-clause utterances. Differential processing has been found for right-branching, cross-serial and center-embedded dependencies (see Christiansen and Chater, 1999b and Gibson, 1998), for subject- and object-relativized subordinate clauses (King and Just, 1991), and more generally the accessibility of noun phrases to relativization (Keenan and Hawkins, 1987). Moreover, differential processing of relative clauses has been shown to co-vary with linguistic experience (MacDonald and Christiansen, 2002; Wells et al., 2008) and frequency of occurrence (Reali and Christiansen, 2007a,b). These studies suggest that probabilistic information influences syntactic processing in intricate ways. Other processing preferences have been observed in relative clause attachment priming (Scheepers, 2003), in the preferred ordering of short noun phrases before longer noun phrases which are modified by relative clauses ('heavy NP shift', Arnold et al., 2000), and in the way main verb/reduced relative clause ambiguities ('garden paths') are resolved in comprehension (MacDonald et al., 1994). Relative clauses are purely syntactic devices to modify NPs and are not part of the argument structure of lexical items or linguistic constructions. Hence, the differential processing of sentences with relative clauses opens a window into human syntactic processing. Processing preferences reveal how syntactic structure is built in the human language processor during comprehension, how meaning is grammatically encoded during production, and shed light on both the nature of the syntactic representations involved and the cognitive architecture maintaining and using these representations.

Relative clauses pose many challenges for theories of processing *and* acquisition. The questions how meaning is mapped onto sentences and how language is learned from input have been studied by different branches of psycholinguistics. The Dual-path model, on the other hand, is built on the assumption that learning and processing are inseparable, that there exists an intimate relationship between linguistic input, syntactic knowledge and the processing capacities of the human language system. The model's processor is the very locus of its syntactic knowledge and this knowledge is shaped through linguistic experience. In the framework of the Dual-path model, the language system is not conceived as a fixed device which constrains learning and processing but as a mechanism which is itself altered and adapted through learning and processing. Since this model is sensitive to the distributional properties of its linguistic environment, it provides an ideal platform to investigate the complex interactions between input, learning and processing of relative clause constructions. In order to utilize the model in this endeavor it is first of all necessary to accommodate its architecture for the processing of multi-clause utterances. This will be the primary task in the current chapter.

The Dual-path model learns from exposure to message-sentence pairs. In the beginning, the model receives a meaning representation as input and incrementally predicts a sentence form suitable to express this message. During prediction, the model's output is compared word-by-word with the intended utterance and the model receives feedback when mismatches occur, which alters the strength of synaptic connections between neurons. In this way, the model's state of knowledge is gradually adjusted until it converges on a stable state which, ideally, represents the target grammar. Thus, representations of syntactic knowledge are the outcome of learning a meaning-to-form mapping for the training language. The model acts as an incremental transducer which casts the conceptual structure of its semantic input into a syntactic string of words. In this process, the model accomplishes a number of subtasks which are instrumental to achieving its learning goal. For instance, it learns the meaning of lexical items in the training language and induces word categories based on statistical regularities in the input. Also, the model learns to appropriately sequence thematic roles in basic constructions and builds representations of syntactic frames for these constructions.[1] Compared with single-clause

[1] See Chang (2002), Chang et al. (2006) and Chapter 5.

structures, sentences with relative clauses complicate the meaning-to-form mapping the model has to learn considerably. It must produce clauses in the right order and respect their integrity. It must identify the relative clause attachment and gapping site, establish the co-reference of constituents in different clauses, omit the relativized constituent in the surface form, and correctly resume superordinate clauses once an embedding is complete. In addition, relative clauses create many alternative forms of expressing identical propositions which renders the meaning-to-form mapping more complex and the message input to the model less distinct. Both these factors might prevent the model from learning the target language to a satisfactory degree altogether.

In this chapter I seek to identify the requirements on the conceptual structure of the message input which enable the model to learn and generalize relative clause constructions. To learn relative clause constructions, the semantic representations of these constructions must be sufficiently rich to allow the Dual-path model to reliably make the right structural choices in processing. Encoding the meaning of complex sentences in the event semantics is subject to three constraints. The first constraint concerns the absence of temporal order with which the message input is provided. The message is given to the model in its entirety at the start of each production episode. All semantic information is present from the beginning and remains active and unaltered until a sentence is complete. The message is static and the model has to dynamically map this message onto a sequence of words. With complete message input the model has to solve a serial-order problem (what to produce next) and a timing problem (when to produce what). To solve these problems, it must figure out when to use which chunk of semantic information in the message. In case of single-clause sentences the model could achieve this task because the conceptual structure of the message corresponded systematically with the syntactic structure of the sentence (see Chang, 2002; Chang et al., 2006). For multi-clause utterances, one might consider providing the model with semantic information in a piecemeal fashion, e.g., separately for main and subordinate clause with a time delay. But this would endow the model with sequential guidance which might not be available to human learners in acquisition or speakers in grammatical encoding.[2]

The second constraint pertains to the spatial organization of the event semantics component of the input message. I aimed at semantic representations which characterize complex events by a linear pattern of activation. Although relations between message elements were encoded by means of neural activation in this pattern, representations did not have a hierarchical structure in a spatial sense (such as, e.g., different layers of features and connections between features in different layers). The model was supposed to assemble the hierarchical structure of complex sentences from activation-based, relational features. Complex events were conceived of as concatenations of more basic events, with semantically salient participants and prominence relations between basic events. All this information was projected onto a flat pattern of activation rather than a hierarchically structured network of semantic feature nodes. The spatial distribution of neural excitation in hierarchically structured representations could give rise

[2] Confer, however, the paragraph on future directions in Section 9.2.3, page 280.

to temporal information in the message input, a property which was ruled out by the first constraint.

A third constraint derives from the learning task itself in which the model is tested on its progress for experienced as well as novel sentences. For learning the experienced fragment of the target language it is conducive if meaning representations are highly distinct. We can expect optimal learning if every construction has an idiosyncratic representation in semantic space, e.g., by using disjoint sets of features. For generalization, on the other hand, idiosyncratic messages are detrimental because novel constructions would be paired with novel meanings not experienced during learning. The model would have no experiential basis for producing structurally novel utterances from semantic representations which share no resemblance with experienced meanings. Messages must therefore be sufficiently close in semantic space to allow similarity-based analogical extension of the production mechanism from trained to novel constructions. Hence, there might be a trade-off between learning and generalization which requires attention in the process of determining a suitable message encoding.

4.2 Artificial language and method

In this section I will briefly describe the artificial language and training conditions I used to find a message representation that enabled the model to learn the syntax of subordination and generalize to novel multi-clause utterances.

4.2.1 Artificial language

This language consisted of templates for linguistic constructions which are basic to human experience, e.g., transitive action and dative transfer. (Goldberg, 1995). These templates contained open argument slots which could be filled by words and inflectional morphemes from the lexicon to create sentences. Table 4.1 lists all basic constructions in the artificial language. By means of relativization, basic constructions were combined to form complex constructions with multiple clauses. In these complex constructions, all combinations of basic constructions were admissible given that the modified head noun matched the relativized element in the subordinate clause in terms of animacy. In this way, a combinatorially complete language with at most one relative clause per sentence was obtained from the constructions of Table 4.1. Some examples of such constructions and their instantiating sentences are shown in Table 4.2. The lexicon from which the argument slots of construction templates were filled contained 72 words and morphemes—two articles (definite and indefinite), 12 animate nouns, 12 inanimate nouns, 23 verbs in four categories (intransitive, transitive, dative and oblique), four auxiliaries (`is`, `are`, `was`, `were`), the continuous form `being`, three prepositions (`by`, `to`, `with`), three inflectional morphemes (`-ing`, `-s`, `-ed`) to mark tense and aspect, a past participle marker (`-par`), the pronoun `that` and an end of sentence marker. In total, the language comprised 131 different constructions which together with this lexicon yielded

Construction type	Example sentence (single-clause)
Animate Intransitive	`the cat was sleep -ing .`
Agent-Patient (active voice)	`the dog is chase -ing the cat .`
Agent-Patient (passive voice)	`a dog is being hit -ed by the teacher .`
Transfer Dative (prepositional)	`the boy give -s a apple to a girl .`
Transfer Dative (double object)	`a nurse show -ed the dog a toy .`
Animate Oblique	`a boy is play -ing with a cat .`

Table 4.1: Basic constructions in the input environment.

roughly 1.03×10^{11} different sentence tokens. Thus, although this language has little variety in its basic construction types, relativization created considerable structural diversity and a large amount of distinct sentence tokens for a data-driven learner to cope with. Clearly, the artificial language is lacking many lexical categories of natural languages such as pronouns, adjectives, adverbs, quantifiers, etc., and I cannot even begin to enumerate the grammatical categories it is lacking. The language is stripped to a structural core of combining basic constructions (and their syntactic alternations) into more complex sentences through relativization. The purpose of using such an impoverished language is to find semantic representations which enable the model to acquire this grammatical device by capturing the relations between events expressed in multi-clause utterances. If these representations are sufficiently general we can easily add linguistic features and constructional variety later on, and tailor the language to more specific learning and generalization tasks.

To train the model, sentences were randomly generated from the templates of Tables 4.1 & 4.2 and then paired with a semantic representation (message) which the model received as input. As described in Chapter 3, each message consisted of concepts, participant roles and event semantics features. The letters X, Y, and Z are placeholders for thematic roles assigned to event participants. These 'semantic variables' were associated with concepts (e.g., BOY, CAT) in the message-lexical system. Combinations of features in the event semantics (e.g., XX, YY, SIMP) encoded the conceptual structure of an intended utterance. As Chang et al. (2006) point out, the XYZ representational scheme does not correspond to any single theory of thematic roles but combines several approaches to meaning. The central Y role is assigned to event participants which are "most saliently changed or moved, or affected by the action", such as subjects of intransitives and obliques, objects of transitives and datives. Participants which cause actions are assigned the X role, such as subjects of transitives and datives. The Z role is assigned to the goal, location or recipient of an action involving movement or transfer, such as dative objects, but also to oblique objects. Table 4.3 shows the role assignment

Construction type	**Example sentence (multi-clause)**
Animate Intransitive (main clause) + Agent-Patient (relative clause)	`the man that kick -ed the dog is run -ing .` (subject-modified—subject-relativized)
Agent-Patient (main clause) + Transfer Dative (relative clause)	`a nurse was hit -ed by the brother that the boy give -s the cake to .` (subject-modified—object-relativized)
⋮	⋮
Transfer Dative (main clause) + Animate Oblique (relative clause)	`a girl present -s the mother that the cat is arrive -ing with a kite .` (object-modified—object-relativized)

Table 4.2: Complex constructions in the input environment.

for each construction in the language from which the input to the model was generated. In the event semantics, each XYZ role corresponded to semantic features whose pattern of activation described the overall event structure, the number and relative prominence of event participants. A simple transitive event (`the dog attack -ed a boy`), for instance, would be represented by activating the agent feature XX, the patient feature YY and the features SIMP and PAST for simple past tense. It is the objective of this chapter to

Construction	Arguments	Action
Animate Intransitive	Y=ANIMAL	sleep, jump, walk, fall, run, arrive
Agent-Patient (active voice)	X=ANIMAL, Y=ANIMAL	push, kick, attack, carry, approach, teach, pat, hit
Agent-Patient (passive voice)	X=ANIMAL, Y=ANIMAL	push, kick, attack, carry, approach, teach, pat, hit
Transfer Dative (prepositional)	X=ANIMAL, Y=THING, Z=ANIMAL	give, throw, show, toss, present, bring
Transfer Dative (double object)	X=ANIMAL, Y=THING, Z=ANIMAL	give, throw, show, toss, present, bring
Animate Oblique	Y=ANIMAL, Z=ANIMAL or THING	jump, walk, run, leave, play, come

Table 4.3: Argument structure of the artificial language. The category ANIMAL comprised humans and pets, the category THING included toys, food and drinks.

analyze different ways of representing the semantic structure of complex constructions by relating atomic events in the event semantics.

4.2.2 Method

For each message type which was compared in the following experiments, the sentence input, training regime and model parameters were identical. Only the semantic representations which the model received as input differed across conditions. The training set consisted of 8.000 simple-clause sentences randomly generated from the six basic constructions in the language and of 2.000 randomly generated sentences containing one relative clause. This input environment is depicted in Figure 4.1. The model was trained on 100.000 sentences in total, effectively cycling through the training set ten times. Sentences from the training set were presented in random order. After every 5.000 training items, the model's learning progress was measured. To do this, the model was tested on 500 simple-clause sentences experienced in training and 500 novel simple-clause sentences which were not experienced in training. This was to ensure that all performance differences the model displayed for relative clause constructions were not due to impaired learning of simple-clause sentences. In addition, after every 5.000 training items, the model was tested on 500 sentences with relative clauses experienced in training, and 500 such sentences which were novel. The performance scores on these sets estimate to what extent the model learned the target language in each condition. The ratio of novel-to-trained scores gives an indication of how well each message representation supports syntactic generalization (see Section 4.3.9). Model performance was measured in terms of sentence accuracy, which compared produced utterances word-by-word with target utterances. To count as a successful production, the model's utterance had to *perfectly match* the target utterance at each sentence position, not only by grammatical category but also by lexical item. Assessing the model's learning behavior, construction types were not distinguished at this stage.[3]

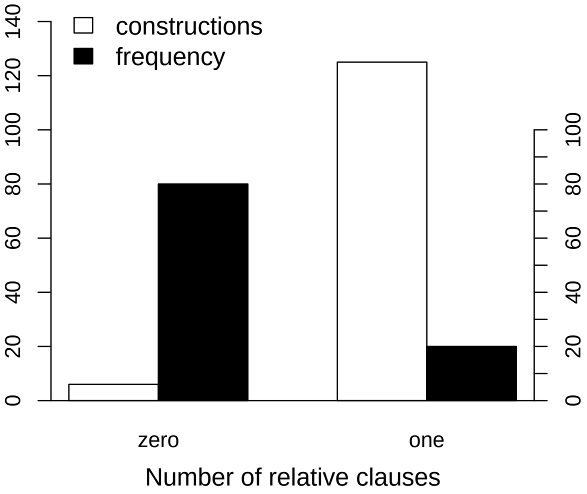


Figure 4.1: Number of constructions in the artificial language and the simple-to-complex ratio in training.

[3] That is to say, tested sentences were not distinguished by the basic constructions they were composed of, not classified into right-branching and center-embedded, subject- and object-relativized constructions,

4.3 Message representation comparison

In this section I will discuss 8 × 2 different ways of encoding the meaning of complex sentences in the Dual-path model's input message. 8 conditions vary the event semantics component of the message and 2 conditions vary the role-to-concept bindings in the model's WHAT-WHERE system. All message types respected the first two constraints from above, they were presented to the model non-dynamically as a linear pattern of activation. The number of compared message types may seem excessive, but the analysis of each condition contributed a piece of insight to the overall puzzle of finding a semantic encoding suitable for the learning and generalization of complex sentence structure. By the end of this chapter I will have identified such a message type and continue using it throughout this thesis.

4.3.1 Random baseline

In the first condition I examined, each construction was assigned an idiosyncratic, holistic meaning. Each event type was represented by a distinct randomized message which—in case of messages with two propositions—was not a combination of representations of simple events. Each construction from the model's input environment plus a unique tense and aspect combination defined a different event type. For example, a simple transfer dative such as

(1) `the cat show -s the kite to the mother .`

counted as a sentence describing a different event type than the sentence

(2) `the cat was show -ing the kite to the mother .`

despite being instances of the same construction. Hence, there was a total of $6 \times 4 = 24$ simple event types and $131 \times 4 \times 4 = 2096$ complex event types expressible in the artificial language. Once a random meaning was assigned to each event type, this message was kept constant across training and testing. Thus, although sentence *tokens* in testing were novel, the random meaning of the underlying event type was identical to the training condition.

In contrast to message representations I will discuss subsequently, there was no default level of activation in the EVENT SEMANTICS layer. All event features were assigned an activation value uniformly randomized between 0.1 and 1.0. For instance, the event type expressed by the sentence

(3) `the mother was being hit -par by the cat .`

from the actual training corpus was represented by setting the activation of the past tense feature to 0.7, the progressive aspect feature to 0.3, the PATIENT feature to 0.2

or separated by other dimensions of distinction which will become relevant in later chapters.

and the AGENT feature to 0.9.[4] The intra-clausal prominence of event participants was not signalled to the model. Moreover, syntactic alternations such as the transitive active/passive constructions were not semantically wedded by systematically varying the figure-ground relationship of the AGENT and PATIENT features (as described in the previous chapter). Each event type received it's own unique and independent representation in the event semantics; the other message elements—thematic roles and concepts—were linked in the usual way in the WHAT-WHERE system.

Among other things, Figures 4.2 (page 82) and 4.3 (page 83) show the model's performance in this 'randomized condition' on single-clause sentences from the training set and on novel such sentences drawn from the same language, averaged over ten training environments.[5] On both test sets—the familiar and the unfamiliar sentences—the model reached close to 100% sentence accuracy. Thus, the model correctly produced the sentences it experienced in the learning phase and generalized to novel simple sentences which it had not experienced in training. Performance looked very different, however, when the model was tested on the multi-clause fragment of the training set. As shown in Figure 4.4 (page 84) it reached only approximately 15% sentence accuracy. For novel sentences with relative clauses the model even scored less than 10% in the RANDOMIZED MESSAGE condition (see Figure 4.5, page 85).

This limitative result suggests that the meaning-to-form mapping for multi-clause utterances is not learnable if the sentence message is non-combinatorial at the propositional level. Learnability requires that clauses which express the same proposition have the same meaning representation whether they occur in a simple or a multi-clause utterance. This is not so much a claim about the compositionality of meaning, but a claim about the necessity of *semantic persistence*. If every sentence, regardless of its number of clauses, has a holistic meaning, linguistic experience is insufficient for learning even a simple language with one level of embedding.

4.3.2 Simple-event message

The requirement of semantic persistence was satisfied in the SIMPLE-EVENT message. In this condition, the event semantics for clauses in complex sentences was identical with the event semantics for clauses in simple sentences. The basic simple-clause message was introduced in Chapter 3 and the description of the artificial language in Section 4.2 above. Complex sentences in the SIMPLE-EVENT message were represented by concatenating the event semantics of two single-clause sentences. For example, a simple transitive sentence such as

[4] The diagram 4.10 on page 102 depicts this and all other compared messages schematically by their characteristic pattern of activation.

[5] By randomizing the event semantics in the described manner, it could accidentally occur that meaning representations of different constructions were too close in 'semantic space'. This could cause the model to not learn certain constructions. Therefore each of the ten model subjects averaged in Figures 4.2 and 4.3 was equipped with a different randomized event semantics.

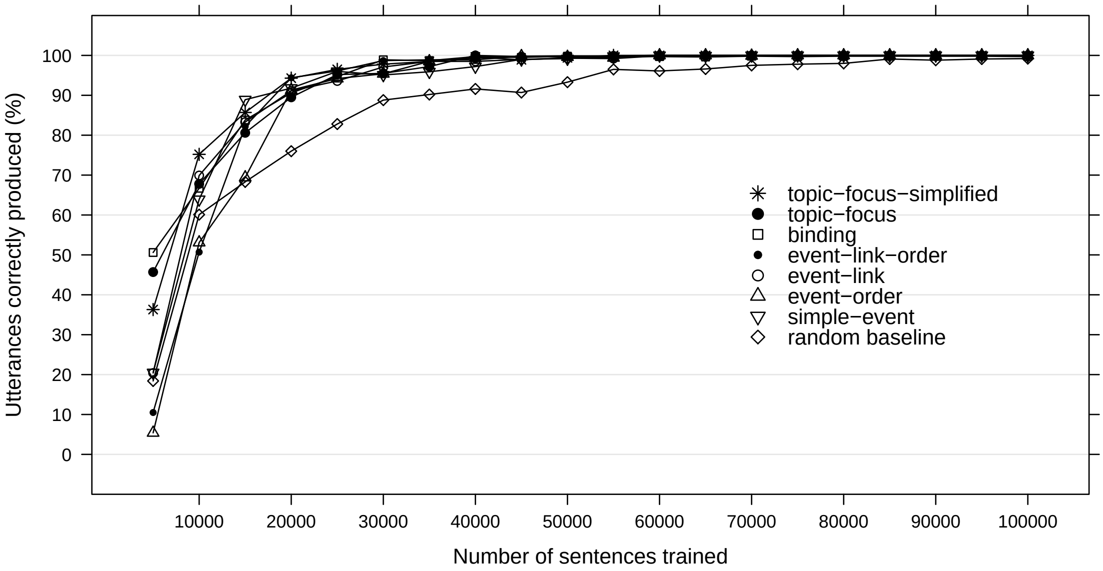


Figure 4.2: Testing on the simple-clause training fragment for all compared message representations.

(4) `the cat chase -s the dog .`

was represented by activating the AGENT, PATIENT, present tense and simple aspect features (XX=YY=PRES=SIMP=1.0) and so was the embedded clause

(5) `...that [the cat] chase -s the dog...`

together with the semantic representation of whichever matrix clause it was combined with. Thus, the model received as input a message for two independent events, but no semantic information about the relation of these events. The SIMPLE-EVENT message was characterized by the inter-sentential persistence of clause meanings and the representation of the intra-clausal prominence of event participants.

Similar to the RANDOM BASELINE, the model rapidly learned to produce correct simple sentences, both trained and novel (Figures 4.2 and 4.3). On trained complex sentences it reached around 50% sentence accuracy, which dropped below 40% for novel complex sentences (Figures 4.4 and 4.5). Because the relative prominence of events was not signalled to the model in its message input, 50% accuracy was the maximum of what could reasonably be expected. Given this limitation of the message, the model performed close to optimal in producing novel multi-clause utterances.

4.3.3 Event-order message

In the EVENT-ORDER message this limitation was removed. One way of thinking about sentences with relative clauses is that they are composed out of two sentences which

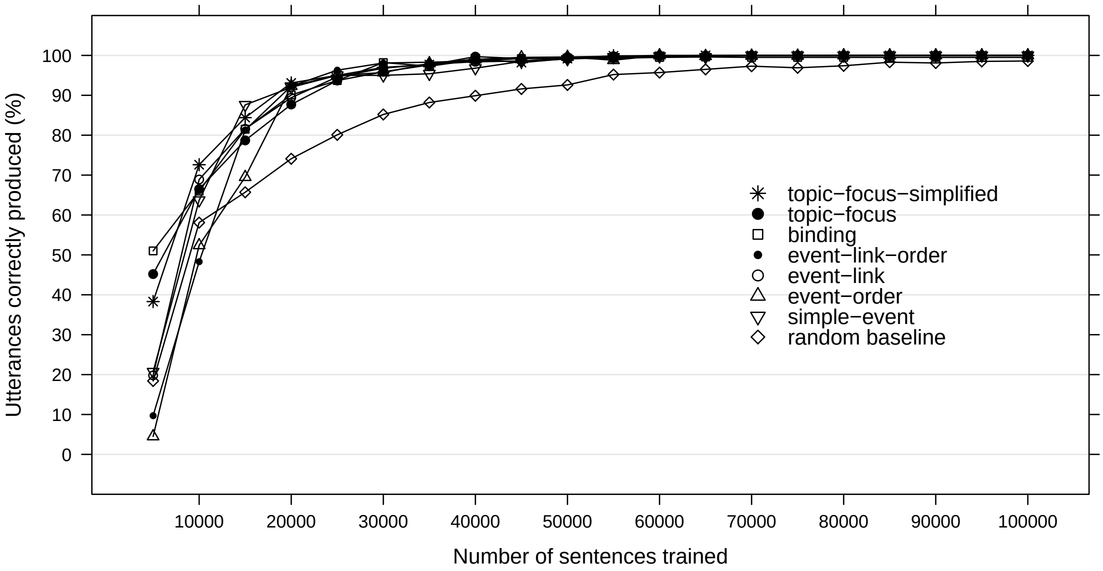


Figure 4.3: Testing on novel simple-clause sentences.

express distinct propositions.

(6) a. `the boy chases the dog .`
b. `the boy runs .`

Both sentences can be related by an anaphoric demonstrative if the referents of the shared NPs are identical in some real world event.

(7) a. `the boy chases the dog .`
b. *`that`* `boy runs .`

The anaphoric demonstrative develops into a relative pronoun as the embedded clause is merged with the main clause.

(8) `the boy that runs chases the dog .`

Parsons (1994, p. 250) speculated that this may be a process by which the restrictive relative clause construction evolved in English, historically. This philological account bears resemblance with the conjoined clause hypothesis of Tavakolian (1981) in language acquisition. According to this hypothesis children process complex sentences as coordinate clausal units, interpret the missing noun phrase as the subject of the relative clause and take it to be co-referential with the subject of the main clause.

Multi-clause sentences describe complex events which are composed out of atomic events (in my artificial language). These atomic events are semantically related since they share an event participant (whose denoting NP is omitted in the relative clause). Shared participants induce relations between events that can be thematic, causal, or

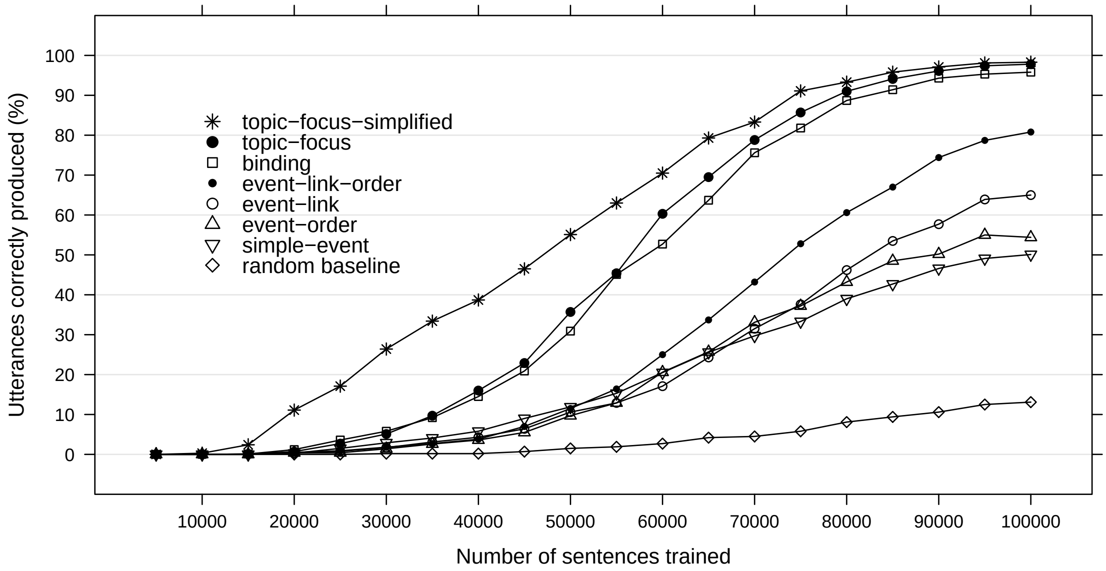


Figure 4.4: Testing on the training fragment containing relative clauses.

temporal in nature. For instance, in the sentence

(9) `the horse that jumps over the fence bit the cow .`

the restrictive relative clause functions as a predicate that identifies the referent of the head noun. The event described in the relative clause specifies a participant in the main clause event and fixes the topic of the sentence. In the sentence

(10) `the man that fell off the bridge died .`

the embedded event is semantically related to the main clause event by causing the latter. In nonrestrictive relative clauses,

(11) `the cat , that chased the dog , is playing in the garden .`

described events are independent, the construction is similar to a conjunction. The relation between the events is predominantly temporal.

The sketched event relations of thematic specification, causal dependency and temporal order semantically structure complex events and it would be desirable to represent such relations in the conceptual structure of the model's message. For the purpose of this chapter, however, it is sufficient to encode a more simplistic *prominence* relation between atomic events. In the literature on discourse and information structure, complex sentences are often analyzed into a foreground and a background information component (Tomlin, 1985; Thompson, 1987). Foreground information is pivotal and central to the discourse whereas background information is peripheral and merely adds material which fleshes out the main events of a narrative. It has been argued that the distinction

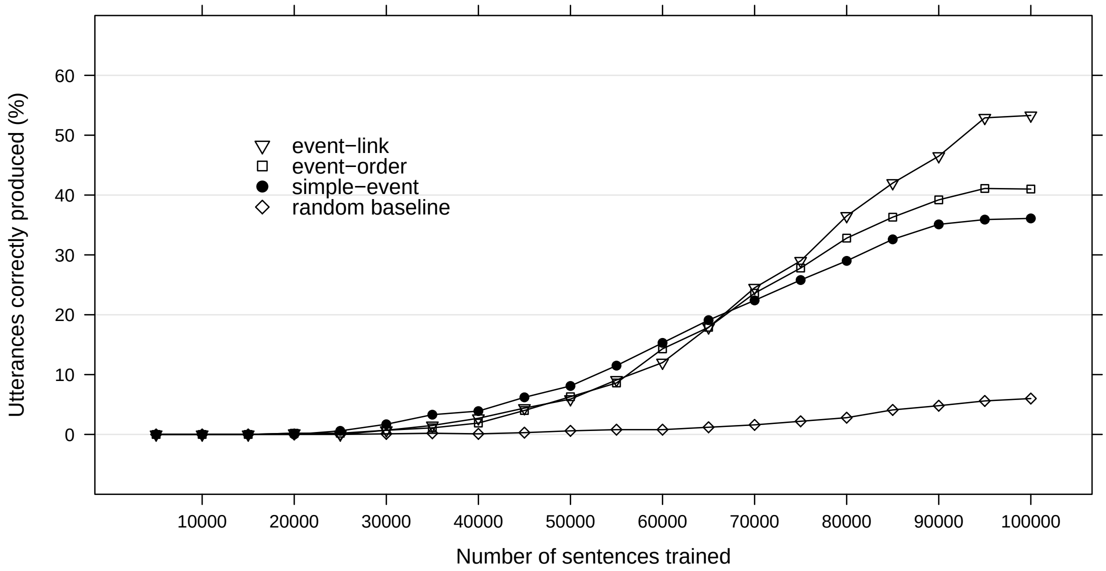


Figure 4.5: Testing on novel sentences containing relative clauses.

between foreground and background information maps onto the distinction between main and subordinate clauses in complex sentences. Main clauses convey crucial information, whereas dependent subordinate clauses supply additional information which is not essential to the narrative discourse (Hopper and Thompson, 1980). Thus, the notions of foreground and background differentiate the semantic content and pragmatic function of clauses in complex sentences.[6] Tomlin (1985), for instance, showed that experimental subjects tend to report events which are important to the narrative by main clauses when describing observed non-verbal action, whereas they express less pivotal information in subordinate clauses.

In addition, a number of studies from the psycholinguistic literature have argued that there are differences in sentence processing between main and subordinate clauses which support the idea that foreground and background information correlate with clause type. Baker and Wagner (1987), for example, showed that readers detect false information more easily in main than in subordinate clauses which suggests that information is more likely to be evaluated for truthfulness when the syntactic structure indicates that it is of central importance rather than logically subordinate and peripheral. The main clause as the central focus of a proposition is more reliably checked against world knowledge than the subordinate clause. Townsend and Bever (1978) demonstrated that the meaning of main clauses is better maintained in memory than the meaning of subordinate clauses whereas verbatim recall is more accurate for subordinate clauses. This indicates that attention is focused on main clause events which are perceived as

[6] Cf. Diessel (2004, p. 44–45). In similar vein, Talmy (2000) proposed that in complex sentences the semantic primitives of *figure* and *ground* characterize main and subordinate clause events, respectively.

more prominent and that semantic information extracted from more salient sentence positions is more available in subsequent processing.[7]

Main and subordinate clauses of complex sentences differ in the types of information they convey and they differ in how accessible this information is in processing. If it is the case that in terms of discourse structure important events are usually expressed by main clauses and less important events by subordinate clauses, this relation of relative prominence between events should be reflected in the processor's representations of the semantic structure of complex sentences. The Dual-path model does not model discourse or the context of utterances, every sentence is presented in isolation and as an autonomous piece of discourse. Thus, unlike a human learner it can not infer prominence relations between events from its learning environment. To compensate for this deficit, the model received semantic information that signalled which atomic event was more pivotal. In this way it could learn to make an informed choice regarding the syntactic structure with which to express complex events. Following the discussion above, I adopted the convention that more prominent events were always expressed by the main clause. In the examples (9)–(11) the prominent events are the temporally posterior, the causal outcome, and the theme of the complex event, but in general prominence is neutral with respect to temporal order or causal direction. The relative prominence of events is anchored in discourse and temporal order or causal direction were therefore not systematically reflected in the syntactic structure of the model's training sentences. For brevity, I will refer to more prominent events expressed by the main clause as the sentence's *theme* and to background events expressed by the relative clause as the sentence's *comment.*[8]

In the EVENT-ORDER message, the relative prominence of events was marked by reducing activation of the tense/aspect features of the comment relative to the level of activation of the tense/aspect features of the theme. The Dual-path model did not use sets of dedicated units in the event semantics to represent different events. An agent feature XX, for example, could signal the presence of an agent in the theme or the comment. If the intended order of events was the default order, the tense/aspect features of both events were switched to 1.0, if the order was inverted, the tense/aspect features of the comment were reduced to 0.5.[9] For example, the event structure of the sentence

(12) `the cat that is sleep -ing chase -ed the dog .`

was represented by the event semantics activation pattern

ES(12) 1XX = 1YY = 1PAST = 1SIMP = 0YY = 1.0, 0PRES = 0PROG = 0.5

[7] Effects in this study varied with the semantic relation between events (being, e.g., causal, temporal, or presuppositional), see also Cooreman and Sanford (1996).

[8] The use of *theme* and *comment* to designate the semantic status of different clauses is not intended to conform with any particular linguistic theory. It is a purely conventional label.

[9] It is also possible to always reduce the comment features but it was easier for the model to focus on the activation state of two nodes instead of comparing the states of 2 × 2 nodes to determine the order of events.

where the 1-features were main clause features and the 0-features are embedded clause features. Hence, in the present condition the message representation employed simple, activation-based means of relating atomic events by placing prominent ones in the foreground and less prominent events in the background. The event-order message not only represented the relative prominence of participants within events but also the relative prominence of atomic events.

Figure 4.5 (page 85) shows that the model tested barely above 40% sentence accuracy for novel complex utterances in the event-order condition. Performance was poor because the message semantically related distinct events in terms of their prominence but contained no information about the event roles of the shared participant. The mere ordering of events did not tell the model which participant in the theme was further specified by the comment. In other words, the event-order message was not sufficiently structured to define a one-one mapping between meanings and sentence forms. Consider the sentence

(13) `the cat chase -ed the dog that is sleep -ing .`

In the event-order message, (13) was represented by the same activation pattern in the event semantics as the previous sentence (12). The relative prominence of events in the message reflected the distinction between main and subordinate clauses, but right-branching and center-embedded constructions are not distinguishable by clause order information alone. A closer look at the model's test sentence output confirmed that incorrect attachment was the major source of the model's production errors. In a word-by-word comparison between target sequences and actual output sequences, the most frequently missed target word was the pronoun `that` and, likewise, the most frequent word erroneously produced (table 4.4). The high number of errors in tense (was/is substitution) also indicates that marking comment and theme on the tense/aspect features was not ideal as they did not fully maintain their designated function.

Ten most frequently missed target words

122 that	92 was	36 is	24 -s	16 to	16 teach	13 being	12 show	11 a

Ten most frequent words erroneously produced

137 that	112 is	35 was	24 -s	14 to	12 blank	9 .	8 a	8 toss	8 -ed

Table 4.4: Most frequent lexical errors in the event-order condition.

Although the model represented some information about the conceptual structure of complex events in its sentence message, it essentially had to guess which two event participants in the theme and comment were co-referential. This condition resembles a communicative situation in which a speaker plans a complex utterance about multiple entities, not knowing the identity of the element which is to be specified or enriched by the relative clause. It seems reasonable to assume that such conceptual content is part of the mental representations of sentence meaning a speaker intends to convey.

The event semantics described in the next section attempts to overcome this deficit by linking events in the message in a more precise fashion.

4.3.4 Event-link message

English relative clauses can be characterized by specifying the syntactic role of the head noun, i.e., the main clause NP which is immediately followed by the relative pronoun, and the syntactic role of the omitted NP inside the relative clause which is coreferential with the head noun. Depending on linguistic theory, these two NPs are frequently called *attachment* and *extraction site* (Levine, 2001), *modified* and *relativized element* (Sag, 1997; Diessel and Tomasello, 2005), or *filler* and *gap* (Wanner and Maratsos, 1978). These terms describe the syntactic properties of constituents in complex sentences and are not usually applied to identify the semantic or pragmatic function of these constituents. Semantically, the modified element is the most prominent sentence constituent because it denotes the unique participant of both events, the theme and the comment. Because restrictive relative clauses are the only modifiers in my artificial language, complex sentences provide strictly more information about the modified element than about any other sentence constituent, irrespective of whether part of this information is backgrounded or pragmatically presupposed. This semantically salient participant I will refer to as the *topic* of the complex sentence in its occurrence in the theme and as the *focus* in its occurrence in the comment. Again, these labels are purely conventional and are required to identify the modified and relativized elements semantically. They do not match the use of 'topic' and 'focus' in any particular theory from the discourse-pragmatics literature.[10]

In the EVENT-LINK message, the model was provided with semantic information about the co-reference of topic and focus. In doing so, I assumed that a language learner can infer the referent of the topic/focus element from the visual environment shared with the speaker. It is the very pragmatic purpose of a restrictive relative clause to single out a discourse referent and establish the topic of a sentence. The presence of multiple referents is therefore a felicity condition for the use of restrictive relative clauses (Córrea, 1995; Kidd, 2003). If there is only one possible referent in a real world event, the relative clause is redundant. If there is a set of possible referents, the relative clause restricts this set and disambiguates the main clause meaning. Language learners might not be sensitive to the pragmatic function of relative clauses but adult speakers are and they might support this function through pointing, demonstration or gesture. Pointing, e.g., serves to direct the hearer's attention to the spatial region of the referent (Marslen-Wilson et al., 1982) and reference resolution is facilitated when speaker and hearer share attentional focus on the same region (Hanna and Tanenhaus, 2004).

[10] According to Lambrecht (1994) the topic is characterized by semantic 'aboutness' and the focus 'enriches the topic semantically'. Since "a relative clause must be *a statement about* its head noun", Kuno (1976) proposed to view the head noun of a relative clause as the topic of that clause (p. 420). The notion of focus is also used by Villiers et al. (1979), Sheldon (1974) and Kidd and Bavin (2002), referring to the syntactic role of the relativized element.

Specifically, then, I assume that the triadic relation between speaker, learner and topic referent can be established in a communicative situation through joint attention and non-linguistic deixis.

Topic/focus information connects events in a different way as theme and comment information. Theme and comment relate atomic events in terms of their relative prominence but do not highlight individual participants, whereas topic and focus identify the co-referential participants of both events. Of course, topic and focus are not canonically associated with a particular syntactic function or thematic role. However, topics and foci can be individuated by their thematic role because no two participants in atomic events occupy the same thematic role. In the EVENT-LINK message, topic/focus information was incorporated by marking the event semantics features corresponding to the thematic role of the joint participant in each event. This was implemented by reducing the default activation of features by a fixed value of 0.3. For instance, if the main clause agent was the sentence topic the 0XX feature in the event semantics was reduced to 0.7. If the relative clause patient was the sentence focus the 1YY feature in the event semantics was equally reduced to 0.7. In this message, the event structure of sentence (12) was represented by the following pattern of activation:

ES(12) 1XX = 0YY = 0.7, 1YY = 1PAST = 1SIMP = 0PRES = 0PROG = 1.0

The event structure of sentence (13), on the other hand, was represented as

ES(13) 1YY = 0YY = 0.7, 1XX = 1PAST = 1SIMP = 0PRES = 0PROG = 1.0

Hence, in contrast to the EVENT-ORDER message from the previous section, the EVENT-LINK message could distinguish both sentences in that it contained information about semantically salient participants.[11]

Because topic and focus map directly to the positions of the head noun and gapped element we could expect strongly improved performance compared with the EVENT-ORDER condition in which attachment errors accounted for a large number of incorrect productions. Sentence accuracy for novel complex test items, however, reached only around 53% (Figure 4.5, page 85). Intuitively, the model's deficits are rooted in the way topic and focus were marked in the EVENT-LINK message. The event semantics contained information about the co-reference of topic and focus but it did not signal to the model *which was which.* In other words, the event features carrying topic and focus content were not distinguished themselves. Consequently, the model could not associate the topic with the theme of the complex event, or the focus with the comment for that matter. Just as in the SIMPLE-CLAUSE message, atomic events were not distinguishable in terms of their relative prominence. If this hypothesis is correct, we should expect that the model generated many errors which reflect main/relative clause confusion. To test this, I examined and classified the actual sentence output of the model for 187 novel input messages (table 4.5). Sentences causing type I errors all were object-relativized and the model produced the relative clause subject NP at sentence onset. If, e.g.,

[11] Again, refer to diagram 4.10 on page 102 for a more perspicuous comparison.

Label	Error type	Number	Cummulative Percentage
I	Initial noun phrase	41	21.8
II	Initial determiner	54	50.8
III	Verb scrambling	18	60.4
IV	Aspect scrambling	10	65.8
V	Attachment	32	82.9
VI	Non-classified	32	100
Total		187	

Table 4.5: Types of production errors in the EVENT-LINK condition.

(14) `the cat chase -ed the dog that a boy is hit -ing .`

was the target sentence, the model started off producing `a boy...` instead of `the cat`. In type II errors, the model initially produced the determiner of the relative clause subject NP together with the main clause subject noun, i.e., `a cat...` for sentence (14). Both types of error indicate that the model had difficulties with clause order, suggesting that it could not determine the relative prominence of events based on the message input. In verb and aspect scrambling—error types III and IV—the model began sentence (14) with the correct main clause NP but then confused either the actions or the temporal flow of the atomic events, as in `the cat hit -ed...` and `the cat is chase...`, respectively. These two error types suggest that the model did not process complex propositions in terms of atomic events, which is witnessed by disregarding the integrity of clausal units expressing these events. A type V error was committed when the model did not identify the topic of the construction and attached the relative clause to the wrong main clause NP. Type VI errors included several forms of incorrect utterances such as wrong determiners in positions other than sentence initial, the wrong choice of syntactic form (e.g., a double object dative instead of an intended prepositional dative), and other nondescript mistakes.

Problems with clause order and clausal integrity in type I–IV errors accounted for nearly two thirds of all production errors in the EVENT-LINK condition. These errors can be attributed to the message input which only signalled the co-reference of topic and focus but otherwise put both atomic events on a par. Next, I combined the EVENT-LINK features of the message with the EVENT-ORDER features from the previous message. The resulting event semantics marked topic and focus of the complex event as well as the relative prominence of atomic events. Given the model's performance and the distribution of characteristic errors in the current condition it can be expected that the model reaches above 80% sentence accuracy when both sets of semantic features are present.

4.3.5 Event-order-link message

The EVENT-ORDER-LINK message combined features for encoding the relative prominence of thematic roles within clauses, the relative prominence of events within a sentence, and the semantic prominence of topic and focus. It was the conjunction of the previous two message types. Consider the sentence

(15) `the cat that a man was hit -par by is show -ing the toy to a girl .`

The event semantics of this sentence in the EVENT-ORDER-LINK message is depicted in Figure 4.6. The 0PRES and 0PROG features marked the present progressive of the theme whereas the 1PAST and 1SIMP encoded the simple past of the comment.

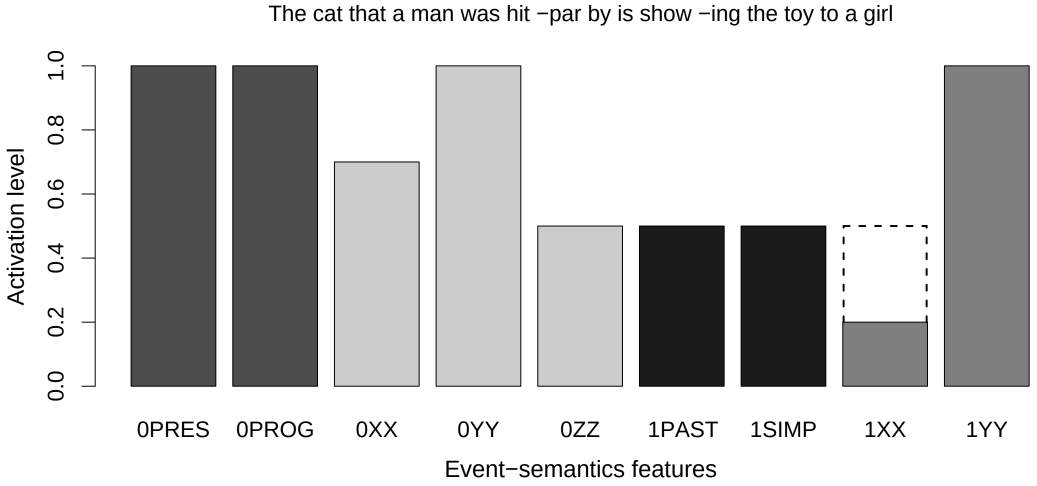


Figure 4.6: Event-semantics in the EVENT-ORDER-LINK condition.

The theme/comment contrast was marked by the activation difference between 0PRES/0PROG and 1PAST/1SIMP, respectively. The intended prepositional dative main clause was represented by reducing the activation of the recipient feature 0ZZ relative to the agent and theme features 0XX and 0YY. The 0XX feature was reduced to mark the agent as the topic of the main clause. Passive voice in the relative clause was represented by reducing the activation of the agent feature 1XX relative to the patient feature 1YY to the level indicated by the dashed bar. Because the agent of the relative clause was also the comment's focus, co-referential with the main clause topic, the agent feature 1XX was reduced a second time to the level indicated by the solid bar.

In contrast to all previous kinds of event semantics, the EVENT-ORDER-LINK message uniquely specified each construction in the language. The mapping between meaning representations and sentence structures was bijective; there were no ambiguities and no two ways of expressing the same message. Arguably, the two sentences

(16) a. `the girl that a boy chase -ed leave -s with a dog .`

b. `a boy chase -ed the girl that leave -s with a dog .`

express the same proposition, but they were assigned a different EVENT-ORDER-LINK message. The difference lies in the salience of atomic events. Because topic and focus were marked as co-referential but were not distinguished semantically, the difference between (16-a) and (16-b) could be represented by simply flipping the relative prominence of events described by main and embedded clause. For the sentence of Figure 4.6 this was realized by fully activating the 1PAST and 1SIMP features and leaving everything else constant. The resulting message mapped onto the sentence

(17) `a man was hit -par by the cat that is show -ing the toy to a girl .`

which has the center-embedded relative clause of (15) as its main clause and the main clause of (15) as its right-branching relative clause. In order to enable this clause alternation in the model, sets of semantic features had to be multi-purpose. Thematic features which represented the prominence of roles in the main clause could also encode the role order in the relative clause and vice versa. Only with this flexibility in the message could the inversion of event prominence bias the model towards producing an alternative structure to express the same proposition.

With the EVENT-ORDER-LINK message, the model learned the simple clause fragment of the artificial language as usual (Figure 4.3, page 83) and reached around 70% sentence accuracy for novel complex sentences (Figure 4.7). Thus, the performance was slightly lower than projected. I examined the production output of a model subject whose accuracy score (71.2%) was closest to the mean over all model subjects. Out of

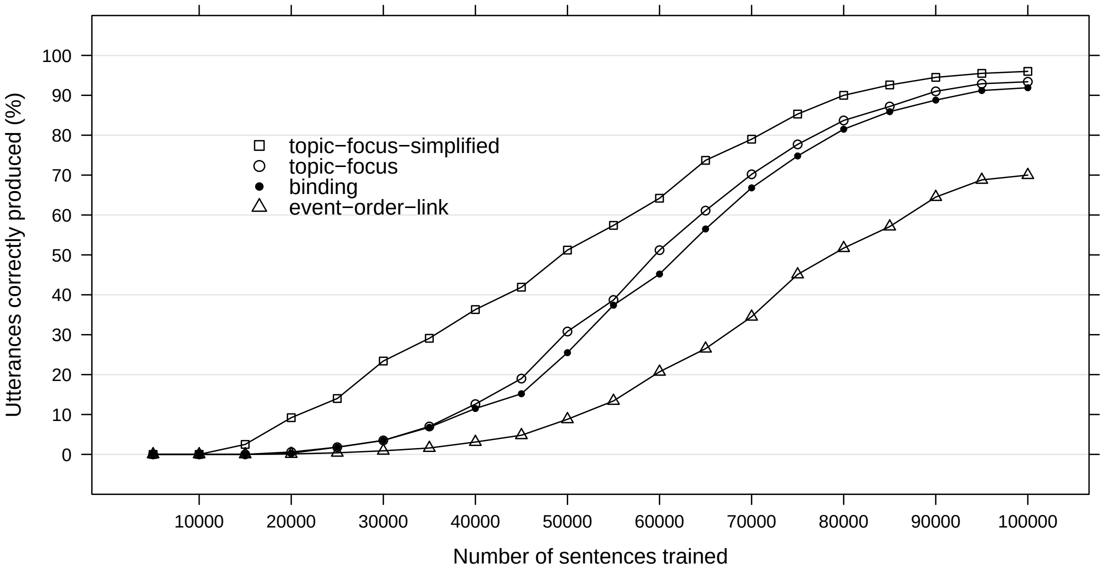


Figure 4.7: Testing on novel sentences with relative clauses.

144 total errors, only 3 (2.1%) involved mistakes which are characteristic of the clause order problems described in the previous paragraph on the EVENT-LINK message. For 99.4% of all tested sentences the model correctly started out producing the clause which expressed the theme; the chosen encoding of the prominence of atomic events was very effective. The remaining errors fell into a variety of categories (wrong determiners, aspect or tense, attachment errors, etc.). The two most common errors, however, both involved the specific way in which topic and focus were marked in the event semantics. First, reducing the activation of message features which corresponded to topic and focus caused interference with the encoding of syntactic alternations. For example, instead of the subject-relativized transitive relative clause

(18) `...with a man that pat -ed the boy .`

the model produced an object-relativized embedding

(19) `...with a man that the boy....`

The intended structure was encoded by reducing the activation of the agent feature in the comment—the focus role—to a value of 0.7. The actually produced relative clause onset, however, suggests that the model misinterpreted the focus information as signalling a passive construction (in which case the same feature got reduced to a value of 0.5). Marking the focus on the role features interfered with the encoding of the passive transitive, because the model could not reliably distinguish activation patterns for different structures. Similarly, the model misinterpreted focus information in a double-object dative comment as signalling a prepositional dative construction.

Reductions of activation in the event semantics carried information. In the EVENT-ORDER-LINK message, reductions encoded the semantics of topic/focus and syntactic alternations. Both types of information interfered and caused the model to produce unintended sentences. It seemed plausible that many errors were due to the small difference in activation between topic/focus and alternation encoding. The model might have been more sensitive to different types of information if activation differences had been more distinct. However, this was not the case, as the second dominant error type in the EVENT-ORDER-LINK condition illustrated. It occurred frequently when there was a double reduction on one event role to signal two different message aspects. For instance, instead of the object-modified passive main clause

(20) `a cat is being push -par by the girl that...`

the model began producing a subject-modified main clause `a cat that....` The intended construction (20) was encoded by first reducing the agent feature to a value of 0.5 to signal passive voice plus another reduction to 0.3 to signal the main clause topic. Thus, the passive marking was obscured by the additional topic marking and the model interpreted the double reduction on the agent feature as attachment information for an active transitive structure. One semantic feature (focus) was masking another (inverse role order for passive voice) in this message. The structure that the model actually produced would have been encoded by a single reduction of the AGENT feature to a value of 0.7. Hence, the activation difference between the message patterns for target and actual sentence, was 0.4 which is quite large. Nonetheless, the model could not distinguish the two messages, i.e., passive object-modified versus active subject-modified transitive. This suggests that the model was not sensitive to the absolute level of activation of a message feature but rather compared levels of activation for different features and opted for the simplest structural choice which was consistent with the activation difference.

To summarize, the EVENT-ORDER-LINK message had two shortcomings. Differential activation could encode syntactic alternations and topic/focus information. Although numerically distinct, the corresponding patterns created ambiguities for the model which caused production errors. Likewise, when one role feature was involved in the encoding of two message properties—such as passivizing the sentence topic—the overlay of information caused incorrect utterances. Both defects will be dealt with in the meaning representations described in the following three subsections. The current condition also showed that message features were not additive, i.e., the model's performance for combining features from two message types (EVENT-LINK and EVENT-ORDER) was not equal to the added performance for both feature sets individually. Combining features can be suboptimal if the message becomes too complicated. Simplification will therefore be aimed at in the subsequent condition, while retaining the unequivocal message-sentence mapping of the EVENT-ORDER-LINK message.

4.3.6 Binding message

In the previous condition, semantic information was encoded relationally and one event feature could be involved signalling two aspects of the message. This was one reason why the model did not learn the complete language from its message-sentence input pairs. To avoid this complication, I detached the two sets of features which encoded the relative prominence of participants within atomic events and the topic/focus and theme/comment relations between these events. The co-reference of topic and focus was no longer signalled by activation differences on role features, but by dedicated nodes in the event semantics. The simplest way to implement this was to utilize special binding features which linked the topic and focus of a complex event. For each pair of event roles that could function as topic and focus, there was a special feature node in the event semantics. This feature was not construction-specific. If the agent of the theme was the topic and co-referential with the patient focus of the comment, a feature was activated which represented this topic/focus relation between participants in the two events, regardless of the particular construction in which the agent and patient roles occurred. Together with the semantic information in the atomic events, this feature uniquely characterized the intended construction. Consider e.g.,

(21) `the cat chase -ed the dog that bite -s a man .`

The event semantics for this sentence in the BINDING message consisted of the SIMPLE-EVENT message for atomic events plus a single feature which marked the patient of chasing (dog) as the topic *and* the agent of biting (dog) as the focus. This semantic feature was denoted by 0Y1X to indicate that the intended sentence for this message had a relative clause attached to the object (=patient) of the main clause, with the subject (=agent) of the subordinate clause relativized. Since there was clause alternation in the input language (and hence no fixed set of event semantics nodes for each clause type), the feature 0Y1X was distinct from the feature 1Y0X. With this semantic distinction the need to signal the prominence of events to the model became superfluous. If 0Y1X was activated the model first had to sequence all event participants indexed with a '0' in the main clause (until the topic was produced) and then all event participants indexed with a '1'. The intended order of role sequencing was reversed if 1Y0X was activated. As a consequence, the BINDING message was particularly simple and parsimonious. Apart from representing the intra-clausal prominence of roles the message employed a single feature to encode the topic/focus binding *and* the theme/comment relationship with only one additional node. There was no need to mark the relative prominence of events on the tense/aspect features (as e.g., in the EVENT-ORDER message). The semantic encoding was non-relational because binding features were autonomous and switched either on or off.

Figure 4.7 (page 92) shows that the model tested above 90% sentence accuracy on novel relative clauses with the BINDING message. Around 78% of the remaining errors involved either wrong tense/aspect, or an incorrect verb or determiner at some position in the sentence. Thus, the majority of production errors was lexical in nature, not structural (e.g., wrong clause order and/or attachment), which can be viewed as an adequacy criterion for a model of complex sentence production. Notwithstanding, the BINDING message was ill-suited for a number of reasons. The downside of parsimony per construction was a rapid inflation of features (and hence EVENT SEMANTICS-layer nodes) as the language got more complex. The number of required binding nodes (BN) depended on the maximum number of roles per clause (RC) and the depth of embedding permitted by the language (DE), and can be calculated as $\text{BN} = \text{RC}^2 \times \text{DE} \times (\text{DE} + 1)$. For example, in the artificial language employed here, there was a maximum of three roles per clause (in the dative construction) and at most one relative clause per sentence. Hence 18 binding nodes were necessary to implement the BINDING message. When the language permitted four embeddings (Chapter 6), the number of required binding nodes was already 180. Large numbers of feature nodes are not immediately problematic, architecturally. Due to the theme/comment-specificity of the binding nodes, however, the model needs more and more training as the input language becomes more complex and the depth of embedding increases. If the model has not been exposed to a training sentence with agent topic and patient focus for a particular combination of events, the corresponding binding node is not trained and the model would not produce such a sentence correctly. For example, training the 0X1Y feature did not enable the model to produce this binding type at a deeper level of embedding. In other words, with the BINDING message the model cannot be expected to generalize topic/focus combinations to novel constructions. To compensate for this inability, the amount of training would have to be increased with the number of embeddings when in fact the exposure to complex sentences decreases with the number of embeddings in realistic learning environments. A second drawback of the BINDING message, which is a consequence of the first, is that the model does not acquire an abstract notion of relative clause topic and focus. Semantic features for topic/focus marking are idiosyncratic and do not recombine. In the next message I tested, this will be remedied. Finally, the BINDING message proved inadequate with respect to a semantic property which will be discussed in Subsection 4.4 below.

4.3.7 Topic-focus message

The BINDING message showed that unambiguous topic/focus information was crucial for the model's high accuracy on novel multi-clause utterances. It was useful to equip the model with separate features to signal topic/focus co-reference instead of marking the corresponding event roles (as in the EVENT-ORDER-LINK message). In the latter condition, topic and focus were symmetric in that both were represented by a reduction of the default level of activation. The BINDING message, on the other hand, was asymmetric because there were two distinct binding features for each pair of roles, depending on which was the topic and which was the focus, respectively. In the TOPIC-FOCUS message I retained the feature separation and the asymmetry of the BINDING message. However, instead of having one feature signal co-reference, there were two separate features, one for the topic and one for the focus. Binding was then encoded by concurrent activation of those features. For each atomic event in the message, there was a set of features which could represent the topic and focus of the complex event. These features were distinct from the event roles which indicated the relative prominence of participants. For example, suppose the target sentence had an active transitive, subject-modified main clause. Then the 0XX feature in the event semantics was activated and there was a separate feature 0XXT in the message to signal that the agent was also the topic of this event. Similarly, there was a feature 1YYF to indicate, e.g., that the patient was the focus of the event in the relative clause of the intended construction. This representation was asymmetric because sets of topic and focus features were distinct. The model had to learn that the topic feature always belonged to the message theme and the focus feature always belonged to the message comment. The set of event features containing the active topic informed the model about the semantic structure of the main clause, the set of event features containing the active focus informed the model about the semantic structure of the relative clause. Therefore it was unnecessary to specially mark the relative prominence of events in the TOPIC-FOCUS condition.

The resulting message was combinatorial. A topic feature could figure in the semantic representation of any construction and combine with any other focus feature. Hence, the binding of event participants could be encoded using only $\mathrm{BN} = 2 \times \mathrm{RC} \times (\mathrm{DE} + 1)$ nodes (e.g., $\mathrm{BN} = 30$ for a language with up to four relative clauses). More importantly, this message potentially enabled the model to generalize the input fragment of the language to constructions which were not encountered in training. It was hoped that the model could acquire abstract notions of topic, focus and co-reference and transfer this knowledge from experiencing specific instances of multi-clause sentences to correctly producing tokens of entirely novel constructions. Whether this was indeed the case will be examined in detail in Chapter 6 on structural generalization.

The TOPIC-FOCUS message allowed minimal semantic changes to cause structurally distinct sentence output which expressed the same proposition. For instance, the two sentences

(22) a. `the man kicks the dog that sleeps .`
b. `the dog that the man kicks sleeps .`

were assigned identical semantic representations except that in (22-a) the sentence topic was placed in the event that involved a man kicking, whereas in (22-b) it was placed in the event which involved a dog sleeping. In this way the model represented which was the foregrounded and which was the backgrounded event and was biased to select the appropriate sentence structure to convey the same proposition with different emphasis on atomic events.

Turning to the analysis of performance, the model reached 93.4% sentence accuracy for novel relative clauses (Figure 4.7, page 92). I examined the production errors of the model subject least deviant from the average score. At epoch 100.000 this model produced 26 incorrect sentences in total (out of 500 tested). None of these errors involved wrong attachment or confusion of main/relative clause. 22 (or 84,6%) of the errors were directly related to syntactic alternations in the language, e.g., the model produced a prepositional dative instead of a double object dative or it produced an active transitive instead of a passive transitive relative clause. This means that most residual errors in the TOPIC-FOCUS condition were clause-internal in nature and did not result from misrepresenting the hierarchical organization of multi-clause utterances. The model had more difficulties ordering thematic roles within clauses, than mapping theme/comment and topic/focus information onto the right construction type. With respect to the current learning task, the TOPIC-FOCUS message was therefore suitable for the processing of complex sentences. Secondly, 16 errors (61.5%) occurred in the relative clause and 10 (38.5%) in the main clause, i.e., the model had more difficulty to successfully produce a relative clause than a main clause. For artificial languages with several levels of embedding we can therefore expect that the model's error rate should increase with clause depth. This will also be tested in Chapter 6.

The non-uniform distribution of errors over clause types suggests that the model is sensitive to the hierarchical structure of sentences and is not processing multi-clause utterances as linear sequences or flat strings of words. This claim is supported by an analysis of error positions within sentences. The average length of incorrect sentences was 16.7 words (sd 1.78), the average error position was 12.3 words (sd 3.79). Thus errors occurred late in the sentence, after nearly 75% of the utterance had been produced correctly. If the model processed sentences as flat strings, based on local transitional probabilities only, we would expect the average error position to be sentence medial, unless transitional probabilities weaken towards the end of the sentence. This might indeed the case since, e.g., all sentences started with a determiner followed by a noun whereas there was more conditional uncertainty later in the sentence when structural choices became available. Furthermore, the sensitivity of the model's sequencing system to immediate word context (SRN working memory) decreases with sentence length because more and more information accumulates in the recurrent buffer. Hence, late errors could be explained by conditional uncertainty plus architectural constraints which make word-to-word transitions more difficult to predict later in the sentence. A closer look at exact error positions, however, revealed that 8 instances occurred at the end of a relative clause (terminal word) and 9 occurred at the continuation of the main clause immediately following a completed relative clause (initial word). Thus, 17 out of 26 errors (65.4%) occurred at the boundary of main and embedded clause. The model had difficulties completing the relative clause and resuming the main clause after the embedding had been produced. This accumulation of errors at clause boundaries indicates that the model is organizing complex sentences into clausal units. It is a claim central to this thesis that the Dual-path model represents the hierarchical structure of multi-clause utterances while at the same time being sensitive to substructure frequencies in linear sequences of words. Throughout Chapters 5–7 I will adduce more evidence for this capacity based on the TOPIC-FOCUS message representation.

4.3.8 Simple topic-focus message

The temporal characterization of events involves two kinds of information, when and how an event happened. Tense is a deictic function which locates an event in time, relative to some point of reference (such as the time of utterance). Aspect characterizes the temporal structure of situations, actions or events as completed, ongoing, etc. English has two basic tenses, past and present, which are morphologically marked by inflectional suffixes (-s, -ed) or by changing the verb stem (run/ran). Aspect is usually subdivided into *lexical* aspect (Aktionsart) and *grammatical* aspect. Lexical aspect classifies situations in terms of temporal features such as being static or dynamic [±dynamic], involving a change of state or location [±telic], or unfolding over time versus occurring in an instant [±durative] (Smith, 1997). Lexical aspect is often considered an inherent temporal property of verbs, although it has been argued that dividing verbs into aspectual classes is ill-conceived because many verbs can be coerced into different Aktionsarten by grammatical constructions (van Lambalgen and Hamm, 2005). Grammatical aspect, on the other hand, classifies utterances in terms of the perspective they convey to the listener and is marked by means of auxiliaries and/or morphemes (e.g., is/was, -ed/-ing). The Dual-path model implements a very simple system of grammatical tense/aspect which can express simple present/past and present/past progressive. Lexical aspect was ignored altogether. Semantically, tense and aspect were represented by the combination of features [±PAST], [±PRES], [±SIMP] and [±PROG] in the event semantics. Thus, it was assumed that tense and aspect characterize the temporal structure of events. Since the EVENT SEMANTICS-layer projected into the HIDDEN-layer, information about tense and aspect of a target utterance could be utilized by both model pathways. The flow of information is schematically shown in Figure 4.8. However, tense and aspect features in the event semantics had no correlate in the conceptual structure of the message in the WHAT-WHERE-system. Hence, to produce auxiliaries and inflectional morphemes at appropriate sentence positions the model had to rely on the sequencing system.

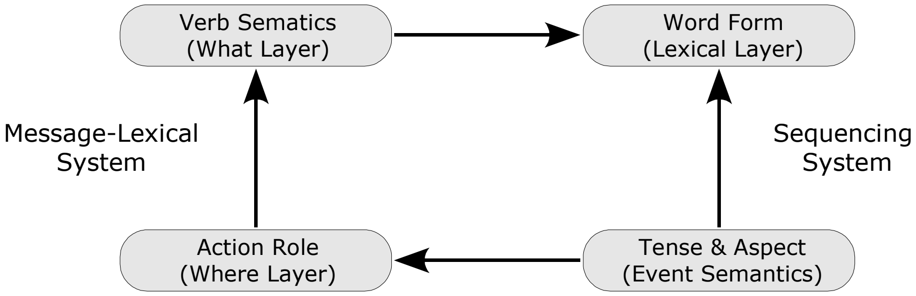


Figure 4.8: Tense and aspect in all previous message conditions.

In all message types studied so far, tense and aspect were conceptualized as general features of events rather than more specific characteristics of the action component of an event. Representationally, tense and aspect features were on a par with features for event participants and were not explicitly related to the action. In the SIMPLE TOPIC-FOCUS message, the semantic representation of tense and aspect was moved from the event semantics into the model's WHAT-WHERE-system, tying it more closely to the action.[12] In the WHAT-layer, nodes were added which represented the concepts of past and present. Likewise, additional nodes represented the concepts of simple and progressive aspect. To encode a specific tense/aspect combination the corresponding nodes in the WHAT-layer were temporarily bound to the action role in the WHERE-layer, along with the usual link to the lexical meaning of the intended verb. Figure 4.9 depicts the way tense and aspect were represented in the SIMPLE TOPIC-FOCUS message. The triple connection of a tense-aspect scheme plus verb meaning to an action role encoded the temporal properties of actions in this message. Note that this implementation is still consistent with the view that aspect is a way of conceptualizing the temporal structure of events. Aspect is treated as an attribute of actions, not as a property of verb semantics. The activation of the action role is determined by the overall event structure and in principle every verb-tense/aspect combination could be associated with an action role.

Placing tense and aspect features in the message-lexical system of the model removed this information from 'sight' of the sequencing system. Tense and aspect were now controlled solely by the activity of the action role in the WHERE-layer. In this way, tense and aspect were separated from syntactic processing in the sequencing pathway.

[12] This was possible because tense and aspect features were not used for theme/comment encoding in the TOPIC-FOCUS message.

This treatment is supported by a study of syntactic priming in language production (Pickering and Branigan, 1998) which showed that the magnitude of priming effects for dative structures was not affected by differences in verb tense or aspect between prime and target, suggesting that the representations involved in syntactic processing are distinct from the representations of verb tense and aspect.

Moving tense and aspect from the event semantics into the WHAT-WHERE-system complicated the mapping from concepts to word forms. The model had to learn the from tense-aspect schemes onto the correct sequence of words in the verb phrase and this mapping was *one-many* (as opposed to *one-one* in all previous conditions). For example, the concept [PRES] for present tense mapped to the lexical item `-s` when the aspect attribute [SIMP] was concurrently active, and to the sequence `is` VERB `-ing` when the aspect attribute [PROG] was active. At the same time it simplified the event semantics. It was not predictable whether this trade-off had a beneficial effect on the model's learning curve. Figure 4.7 (page 92), however, shows that the model learned considerably faster in the SIMPLE TOPIC-FOCUS condition than with the BINDING or TOPIC-FOCUS messages. At epoch 100.000 it reached 97.2% sentence accuracy which was nearly four percent higher than in the TOPIC-FOCUS condition. Of the few remaining errors, 16% involved lexical substitution (wrong article or noun) and 84% involved structural errors in alternations (active/passive and double object/prepositional dative). 80% of all errors occurred in the relative clause, 20% in the main clause. The error profile indicates that the model had no difficulty in learning the mapping from tense/aspect features to complex verb phrases. Better and faster learning indicates that the simplified event semantics facilitated the learning of syntactic frames from the message input.

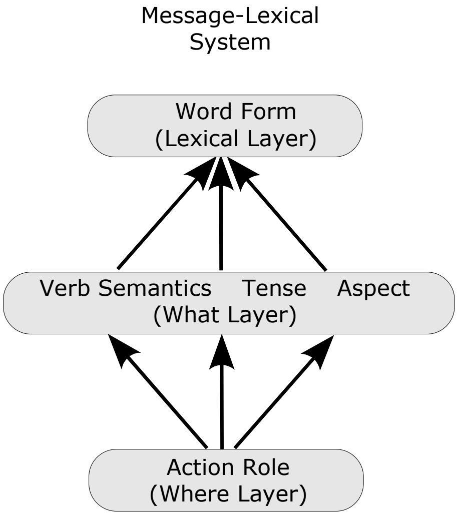

Message-Lexical System
Word Form (Lexical Layer)
Verb Semantics Tense Aspect (What Layer)
Action Role (Where Layer)

Figure 4.9: Tense and aspect in the SIMPLE TOPIC-FOCUS message.

Since the task of learning a target language with relative clauses was the benchmark of this event semantics comparison, the SIMPLE TOPIC-FOCUS message came off as the winner.

4.3.9 Summary message types

All compared messages were adequate for learning the single-clause sentences that were experienced in training (Figure 4.2, page 82) and generalized this knowledge to novel single-clause sentences outside the training set (Figure 4.3, page 83). The random baseline showed that semantic persistence was important for learning multi-clause sentences

(Figure 4.4, page 84). It was a minimal requirement for the message to encode clause meaning invariantly, whether that clause occurred in a simple or complex sentence. The complete array of compared message representations is depicted graphically in terms of activation patterns for the sentence

(23) `the cat that a man was hit -par by is show -ing the toy to a girl .`

in Figure 4.10 on page 102. Only three of these message types proved suitable for learning the multi-clause sentences from the training set and generalized this knowledge to novel complex sentences not experienced during training (Figure 4.7, page 92). These were the BINDING message, the TOPIC-FOCUS message and the SIMPLE TOPIC-FOCUS message. The three models reached >90% sentence accuracy on both these sets of complex utterances.

Performance data from the message comparison was analyzed by a repeated measures ANOVA with the event semantics condition as factor. The dependent variable was the model's sentence accuracy for novel complex utterances measured at epoch 100.000. Not surprisingly, the ANOVA yielded a significant main effect for message type [$F(7,9) = 129.24$, $p < 0.001$]. Post-hoc analyses using the Tukey HSD test indicated that most pairwise comparisons were significant, with the exception of the three conditions which scored the highest. The full comparison matrix is shown in Table 4.6. Accordingly, in terms of performance on novel relative clause sentences the three message types (BINDING, TOPIC-FOCUS, and SIMPLE TOPIC-FOCUS) are indistinguishable. Nonetheless, the TOPIC-FOCUS message is preferable over the BINDING message for rea-

Pairwise message comparison with Tukey HSD

	random baseline	simple event	event order	event link	event order-link	binding	topic-focus	simple topic-focus
random baseline		+	+	+	+	+	+	+
simple event	+		−	+	+	+	+	+
event order	+	−		−	+	+	+	+
event link	+	+	−		+	+	+	+
event order-link	+	+	+	+		+	+	+
binding	+	+	+	+	+		−	−
topic-focus	+	+	+	+	+	−		−
simple topic-focus	+	+	+	+	+	−	−	

Table 4.6: Post-hoc analysis matrix for all event semantics conditions, a + sign marks significant contrasts.

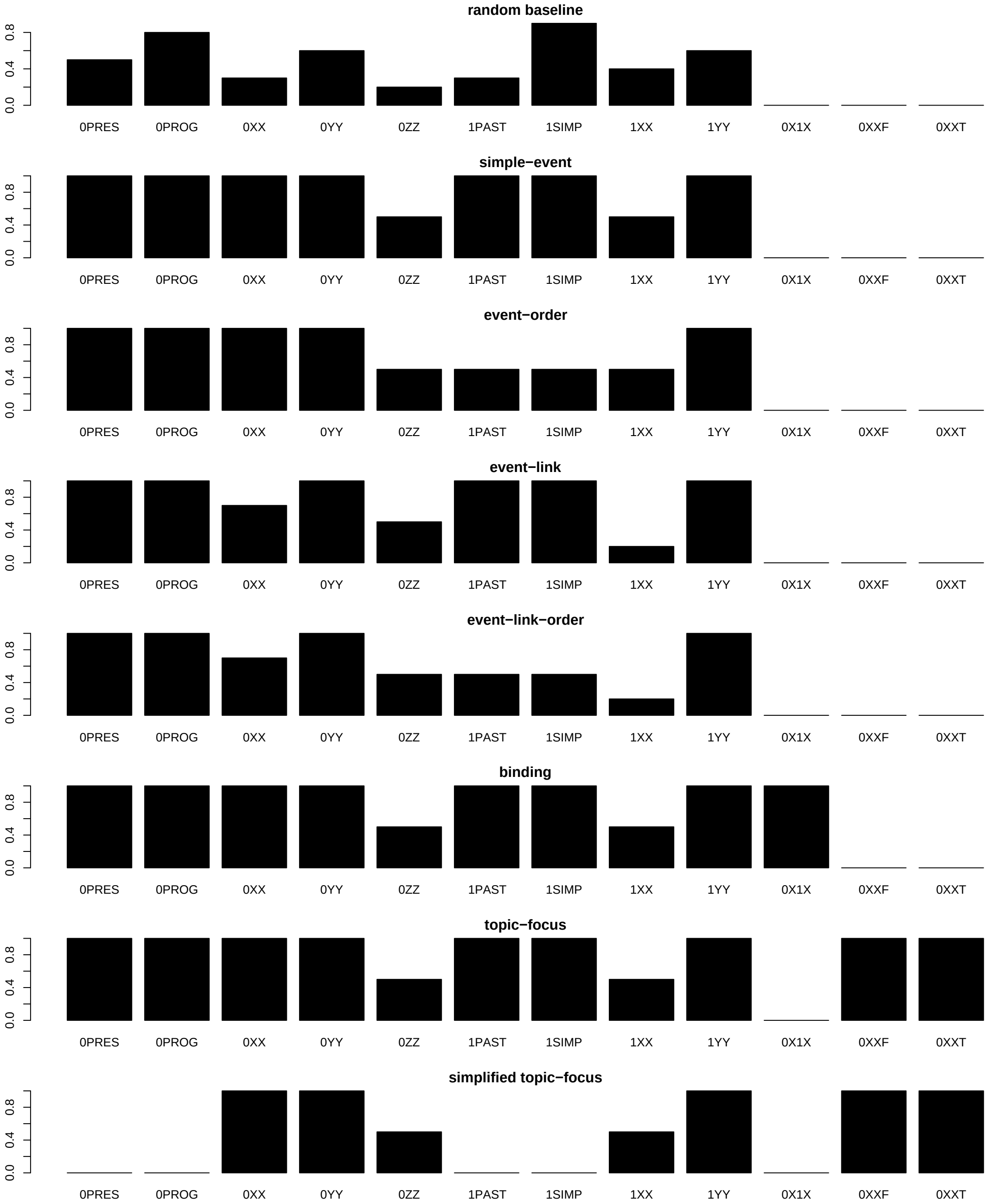


Figure 4.10: Event-semantics activation patterns in comparison for the sentence `the cat that a man was hit -par by is show -ing the toy to a girl`.

sons of parsimony—it used considerably less event semantics nodes to encode sentence meaning. The SIMPLE TOPIC-FOCUS message is preferable over the TOPIC-FOCUS message, because the model learned faster with this meaning representation. The elegance of the SIMPLE TOPIC-FOCUS message compared to other event semantics is visualized in Figure 4.10 (page 102). Models with distinct messages were analyzed in one learning condition. But the parameters of this condition might not be optimal for all messages. The capacity for generalization might therefore better be measured in terms of performance for novel relative to trained sentences. Optimal generalization then occurs if the ratio of these two data points is close to one. This measure of generalization (at the end of training) is plotted in Figure 4.11.[13] With the SIMPLE TOPIC-FOCUS message the model produced novel relative clauses with nearly the same accuracy as trained such sentences. This message is therefore preferable over its competitors with respect to this criterion.[14] Further reasons for preferring the SIMPLE TOPIC-FOCUS message will be given in the next section. All messages suitable for the learning task shared three features which turned out essential for successfully producing complex utterances. First, they all encoded the relative prominence of participants in a systematic and semantically persistent way. The semantic structure of atomic events was represented by activation patterns which were systematically related to a sentence form as, for instance, in the active/passive alternation. These representations were stable across occurrences in different constructions. Consequently, the message was combinatorial at the clausal level. Secondly, all messages marked the topic and focus of complex events. This feature informed the model which participant corresponded to the argument of the relative clause, and which thematic role this participant occupied in the relative clause. The BINDING message employed non-combinatorial, joint topic/focus nodes, the TOPIC-FOCUS messages used clause-specific, separated topic and focus nodes. And third, the message encoded the relative prominence of events in complex propositions, i.e., which atomic event was the theme and which the comment. This information guided the model in producing the cor-

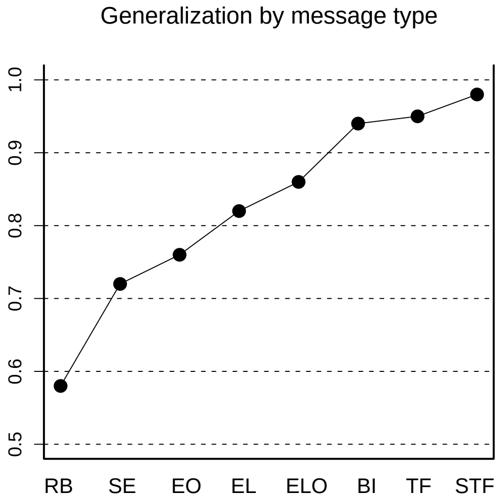


Figure 4.11: Ratio of sentence accuracy for trained to novel complex utterances.

[13] Labels mean: RB = random baseline, SE = simple-event, EO = event-order, EL = event-link, ELO = event-link-order, BI = binding, TF = topic-focus, STF = simple topic-focus.

[14] There was no trade-off between learning and generalization as projected in the introduction. The better the model learned for a message type, the better it generalized. The reason is that, statistically speaking, the model was not tested on novel sentence types but only novel tokens. In later chapters, generalization tasks were more demanding and there such a trade-off did show.

rect order of clauses. The BINDING message represented theme and comment through distinct topic/focus nodes for default and inverted clause order. In the TOPIC-FOCUS messages, the model could infer the theme/comment distinction from the index of the active topic/focus nodes. Apart from these essential traits, I argued that it was important to separate dimensions of information because the model could not trace overlaid message features in multiple reductions of activation (EVENT-ORDER-LINK message versus BINDING message). Furthermore, it was shown that simplifying the message without loss of semantic information can result in a substantial learning speed-up (TOPIC-FOCUS message versus SIMPLE TOPIC-FOCUS message). The distinctive properties of all compared messages are summarized in Table 4.7.[15]

Message Type	System. & Persistent	Theme & Comment	Topic & Focus	Tense & Aspect	Constr. Specific	Coding Conflict
Random baseline	no	no	no	no	yes	no
Simple event	yes	no	no	yes	no	no
Event order	yes	yes	no	yes	no	yes
Event link	yes	no	yes	yes	no	yes
Event order-link	yes	yes	yes	yes	no	yes
Binding	yes	yes	yes	yes	yes	no
Topic-focus	yes	yes	yes	yes	no	no
Simple topic-focus	yes	yes	yes	no	no	no

Table 4.7: All message representations and their specific properties.

4.4 The gapped element

In all message representations I discussed, the relative clause focus had a special status compared to the other event participants. Non-focus participants were anchored in the message twofold. In the WHAT-WHERE system, the thematic role the participant occupied was bound to a concept node in the lexical semantics. In the event semantics a feature corresponding to this role was activated to signal the intended construction (in conjunction with other features). The focus participant of a complex event, on the other hand, was represented only in the event semantics and there was no binding link to the focus concept in the message-lexical system (Figure 4.12). The focus feature in the event semantics was strictly necessary to encode the event type. An example will clarify this point. Suppose, the relative clause focus was the agent of an active transitive sentence. If there was no focus feature in the event semantics, only the patient feature would have

[15] Abbreviations used in Table 4.7: System. = Systematic, Constr. = Construction.

been active in the semantic representation of the relative clause. But this would be compatible with a subject-relativized oblique and a passive transitive. Hence, if the focus feature was absent from the event semantics, the model would not be able to determine the intended construction, nor the clause-internal structure, the target structure would be semantically underdetermined by the message.

The situation was different for the synaptic binding between the focus role in the WHERE-layer and lexical meaning in the WHAT-layer. This message component was not necessary to distinguish all constructions by the message and was omitted for simplicity. It was dispensable because the lexical item associated with the focus position is not overtly produced in English relative clauses. Consequently, there was no need for a dynamic link between the focus role and its filler in the message-lexical system.

I will argue here, however, that for several reasons the gapped element should be present in the meaning representation of a sentence in just the same way as all other event participants.

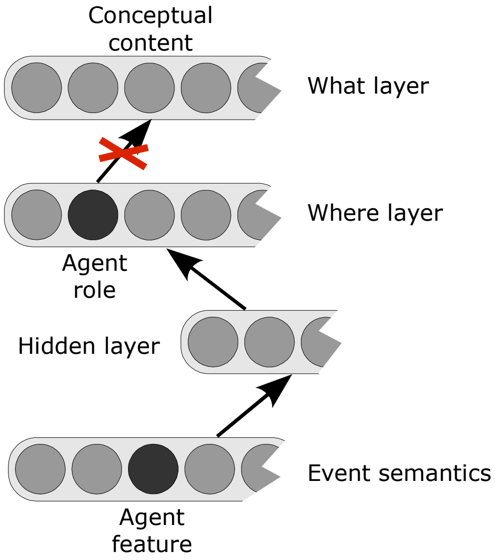


Figure 4.12: The gapped element is represented in the event structure but not linked to conceptual content.

4.4.1 Ambiguities

In English, the relativized NP is not formally expressed inside a restrictive relative clause (*gapping*) and the pronoun *that* carries no information about the syntactic role of the gapped element. In other languages, such as Hebrew, it is possible to indicate the position and role of the relativized NP with a personal pronoun (*pronoun retention*).

(24)	**ha-sarim**	she-ha-nasi	shalax	**otam**	la-mitsraim
	the-ministers	COMP-the-president	sent	**them**	to-Egypt
	'the ministers that the president sent to Egypt'[16]				

Thus Hebrew retains more overt information about the gapped element in the sentence form. In English coordinate structures gapping can cause ambiguity. Consider the sentence

(25) John met Paul yesterday and Ben today.

[16] Example from Max Wheeler, "Relative clauses and the noun phrase accessibility hierarchy." Linguistic typology handout, University of Sussex, 2006.

which has two interpretations, depending on which phrase was omitted

(26) a. John met Paul yesterday and [John met] Ben today. (*conjunction reduction*)
b. John met Paul yesterday and Ben [met Paul] today. (*gapping*)

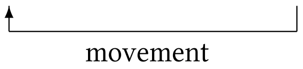
Understanding (25) in either way involves establishing the omitted NP's thematic role and its co-referential element in the first conjunct. The omitted NP is characterized by its 'position' in the argument structure and its conceptual content. It is not merely the agent or patient of the second coordinate clause which is omitted but a *specific* agent or patient that has previously been introduced into the discourse. Comprehension requires to establish the thematic and syntactic role of the gapped element but also its *identity*. Semantically, the gapped element should therefore be treated like any other non-gapped event participant and not be represented as an empty role without filler content.

Ambiguities as in coordinate structures do not occur in unreduced relative clauses (only temporary uncertainty about roles) because the pronoun immediately gives away the identity of the gapped element. However, there is no reason to treat both forms of omission in different ways, the gapped element should be considered a full-fledged message component. This treatment is also more adequate cross-linguistically. Different languages reveal different kinds of information about the gapped element in the sentence surface form—none in English, a resumptive pronoun in Hebrew, gender, number and case (and hence thematic role) marked on German relative pronouns, etc.[17] As the case of Hebrew shows, not all languages simply omit the relativized position but some retain a pronominal marker referring back to the head of the relative clause.[18] This marker is co-referential with the head and has the same conceptual content. For the Dual-path model message this suggests that not only should the gapped element be represented by its role feature in the event semantics but the thematic role itself should be linked to conceptual content in the lexical semantics.

4.4.2 Acquisition

In generative grammar it is common to analyze gaps in relative clauses as the result of some kind of movement. A dislocated constituent is associated with a syntactically dependent empty category which functions like a silent copy of this constituent (see Chomsky, 1995). For instance, in X-bar syntax the gap arises because the relativized element moves from the DP inside the CP to the specifier (SPEC) of the CP and leaves a trace t_i:

(27) [$_{\text{DP}}$ [$_{\text{D}}$ the] [$_{\text{CP}}$ [$_{\text{SPEC}}$ man] [$_{\text{C'}}$ [$_{\text{C}}$ that] [$_{\text{IP}}$ [$_{\text{DP}}$ t_i] [$_{\text{VP}}$ returns the book]]]]]



[17] Languages often have several relativization strategies which retain different kinds of information.
[18] Pronoun retention also occurs in, e.g., Persian, Welsh and Cantonese.

But of course this is merely a theoretical model of relative clause syntax, not a model of the psychological mechanisms of relative clause formation, processing or acquisition. Such a model of syntax does not automatically establish the psychological reality of transformational movement.[19]

Several studies have investigated the psychological reality of traces in sentence comprehension of adults (Swinney et al., 1988; Love and Swinney, 1996; Clahsen and Featherston, 1999) and children (Love, 2007; Roberts et al., 2007) using cross-modal lexical priming. These studies found that semantic information which is linked to the antecedent becomes 'reactivated' at the trace position during on-line comprehension. The parser reconstructs "grammatical and semantic features of the dislocated constituent at a potential gap site by creating a silent syntactic copy of the antecedent" (Roberts et al., 2007, p. 178). This is known as the *trace reactivation hypothesis*. Antecedent priming has also been explained by the *direct association hypothesis* according to which filler-gap dependencies are resolved by reconstructing the verb's argument structure (Pickering and Barry, 1991; Sag and Fodor, 1994). Cross-linguistic evidence from scrambled double-object constructions in German and head-final languages like Japanese in which objects precede verbs, however, is not compatible with the direct association hypothesis (Clahsen and Featherston, 1999, Nakano et al., 2002). Regardless of which is the correct account, mental reactivation effects at gap sites, which have been confirmed by many studies for adults and children, suggest that syntactic gaps play an important role in sentence comprehension.

'Reactivation' of grammatical and semantic features is logically neutral with respect to movement, the 'trace' could also result, for example, from deletion. In developmental psychology it has been hypothesized that early relative clauses merge two simple-clause sentences into a novel constructional unit which expresses a single proposition (cf. Diessel and Tomasello, 2000):

(28) [There's a rabbit]$_1$ that [I'm patting [a rabbit]]$_2$.

The presentational scheme *There's Y* introduces a discourse referent, focuses attention on it, and makes it available for further specification by a relative clause. The object of the transitive scheme *X is VERB -ing Y* gets linked to the presentational clause subject and is deleted from the surface form of the amalgamated construction. No movement or dislocation of constituents is required. In production, the element in the transitive construction which is co-referential with the antecedent becomes phonologically null. According to this view, the acquisition of relative clauses involves the fusion of simpler constructions, and the deletion of multiple occurrences of co-referential elements.[20]

If this mechanism is plausible, the gapped element should be considered a full

[19] How movement rules might be incorporated into a performance model, however, is discussed at length in Jackendoff (2002).

[20] Labelle (1996, p. 68) has suggested that relative clauses are formed by converting a clause into a semantic predicate and co-indexing it with its subject. On her account there is neither movement nor deletion involved.

component of the relative clause's semantic structure, because the complex sentence is formed from two autonomous simple sentences. It is omitted from expression but persists in the conceptual structure of the complex sentence clause due to the combinatorial nature of relative clause formation. In the Dual-path model, binding the role of the relativized element to conceptual content makes this content available for production. Activation of a thematic role at the gap position will activate the lexical meaning of the gapped element which will activate a corresponding word form. Consequently, deletion requires that the model learns to *suppress* the relativized constituent.

4.4.3 Alternations

The artificial language which the model was trained on is sufficiently rich to express complex propositions in a variety of ways. To describe complex events, a number of structural alternatives is available and we would like to be able to bias the model towards selecting one structure over another with only a minimal change in the message. Consider the sentences

(29) a. The man kicks the dog that [the dog] chases the cat.
b. The man kicks the dog that the cat is chased by [the dog].
c. The dog that the man kicks [the dog] chases the cat.
d. The dog that [the dog] is kicked by the man chases the cat.
e. The cat is chased by the dog that the man kicks [the dog].
f. The cat is chased by the dog that [the dog] is kicked by the man.

Arguably, (29-a)–(29-f) express the same proposition. Notice, however, that the gapped instance of *dog* in (29-a), the relative clause focus, becomes the overt main clause topic in (29-c)–(29-f). In these alternations, the embedded clause becomes the main clause (and vice versa) and a right-branching construction becomes a center-embedded construction in (29-c) and (29-d). Furthermore, *the dog* can alternate between agent/patient and subject/object. It was mentioned previously that the message representation of the Dual-path model permitted clause alternation. There were no clause-specific nodes in the WHAT-WHERE-system, or dedicated features in the event semantics. Hence, it was possible to represent the difference between (29-a) and (29-c) simply by bringing the *comment* event to the foreground, making the *theme* event recede in prominence. In the TOPIC-FOCUS message, for example, differential emphasis could be placed on events by inverting the topic/focus relation of the co-referential participants. The model could infer which instance of *dog* should be the attachment and which the gapping site, and thus which instance should get expressed or suppressed, respectively. In order for this alternation bias to work, however, it is necessary that the thematic role node of the gapped element carries conceptual content in just the same way as the thematic role node of the modified element. The role-to-concept bindings in the WHAT-WHERE system must be identical for each of the sentences (29-a) through (29-f) if an instance of *the dog* that is omitted in one alternation becomes expressed in another. Consequently, the focus element should be represented semantically like all expressed event participants. The semantic difference between the six alternations is then reduced to differences in the relative prominence of events and/or event participants (active/passive alternation).

This discussion of how to represent alternations parsimoniously concludes the argumentation for a conceptual link of the focus role to semantic content in the Dual-path model's message.

4.4.4 Performance comparison *gap-link* versus *no gap-link*

I argued that there should be a WHAT-WHERE binding for the gapped element in the message (henceforth: *gap-link*). This *gap-link* might complicate learning and generalization because activating a message element which is available for production must be suppressed. However, the *gap-link* added information to the message, so some message types might profit, others might suffer from it. If for some of the eight candidate messages performance is better with the *gap-link* than without, this would indicate that these candidates implement more adequate meaning representations (within the model's framework).

Figure 4.13 shows all message types tested on the same set of novel complex utterances at epoch 100.000 in both conditions, with and without *gap-link*. A two-sided *t*-test was performed for each pair of messages and significant differences in mean are marked with a diamond. Among the three contenders from the learning comparison, the BINDING message was the only one which scored worse in the *gap-link* condition. Both variants of the TOPIC-FOCUS message, on the other hand, scored higher in the *gap-link* than the *no gap-link* condition. In particular there was a significant effect for linking

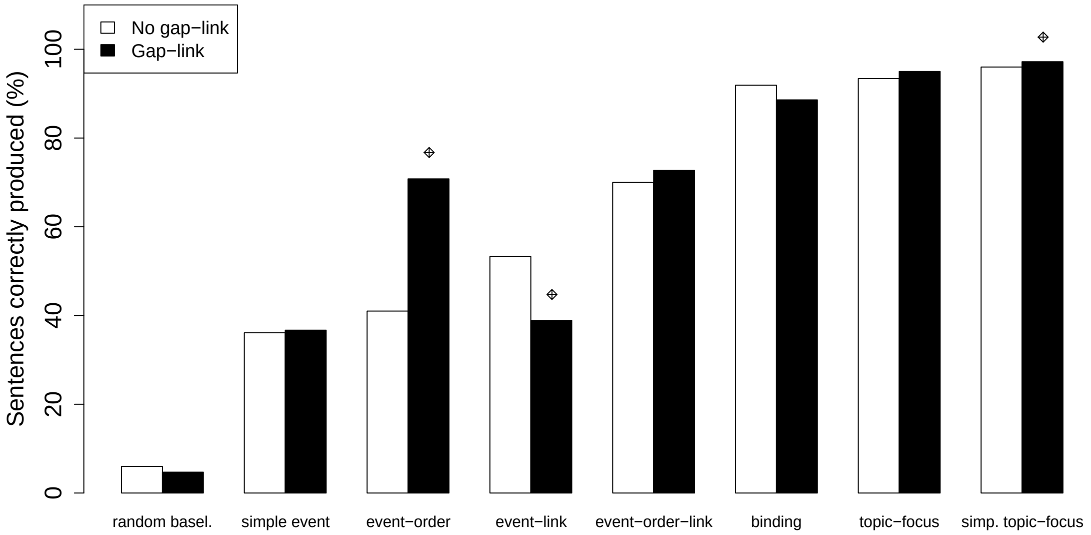


Figure 4.13: Performance comparison for the gapped element linked and disconnected in the message.

the gapped concept in the SIMPLE TOPIC-FOCUS message ($t(9) = 2.34$, $p < 0.05$). This observation makes the SIMPLE TOPIC-FOCUS message the uniquely preferred representation for further simulations. It is also worth pointing out that the EVENT-ORDER message profited most from the *gap-link* in terms of absolute performance gain. With *gap-link* this message type almost reached EVENT-ORDER-LINK message performance, suggesting that the *gap-link* helped the model to establish attachment and gapping sites. The significant drop in sentence accuracy for the EVENT-LINK message with *gap-link*, however, seems to contradict this hypothesis. A closer analysis was required.

I directly compared error patterns of two model subjects from both conditions for both messages. Without *gap-link* there was a total of 261 inaccurate novel complex sentences produced with the EVENT-ORDER message. 225 (or 86%) of these were attachment errors, all of which occurred in center-embedded structures, with 219 of these (97%) involving the sentence initial subject NP. With *gap-link* the number of production errors dropped to 79 which comprised only 6 attachment errors (8%). All these errors involved attachment on the recipient of a double-object dative construction. The largest generic group of errors occurred in embedded passives where the model was confused about the thematic role of the gapped element (agent/patient, 30 times, or 38%). Remaining errors involved wrong aspect, determiners, or some constituent scrambling in the dative alternation. The nearly complete absence of attachment errors in the *gap-link* condition suggests that the model (with EVENT-ORDER message) was able to utilize the fact that two distinct roles were linked to the same concept in the WHAT-WHERE-system in order to determine the special status of the topic/focus element in the message. Together with information about the relative prominence of events this enabled the model to project the attachment and gapping site: due to the EVENT-ORDER information in the message, the model was able to start out producing the intended main clause. At step n in the output sequence, the model produced the main clause topic being ignorant about this constituent's special status as the argument of a relative clause. Due to feedback, the topic concept was activated in the CWHAT-layer at step $n{+}1$ (provided the model had already learned the correct word-to-meaning mapping in the comprehension direction). Since there were two role-to-concept links in the WHAT-WHERE system in the *gap-link* condition (one such link for the topic and one for the focus element), there were two concept-to-role links in the inverse CWHAT-CWHERE-system. Consequently, the topic concept activated two thematic roles in the CWHERE-layer, the roles of the message topic and focus, respectively. This double activation occurred at the sentence position where *that* was to be produced next and it occurred at no other constituent. Because the CWHERE-layer fed into the HIDDEN-layer, the model could utilize this information via its sequencing pathway and attach the relative clause to the target element. Without the *gap-link*, on the other hand, the role of the focus element did not get activated following the production of the topic. Hence the model had no strategy to make an informed structural choice at the attachment site. This advantage of the *gap-link* model explains the large difference in performance for the EVENT-ORDER message.

With the EVENT-LINK message, performance declined when the *gap-link* was added to the message. Again, two model subjects were compared for their specific error profile in both conditions. An attachment error occurred whenever the pronoun *that* was produced in the wrong sequential position. 105 out of a total of 273 incorrect utterances (38%) contained attachment errors when there was *no gap-link*, in contrast to only 5 out of 301 test errors (2%) when the *gap-link* was present. Thus, as with the EVENT-ORDER message, the *gap-link* facilitated structural selection. This is attested by the fact that despite the sentence accuracy went down from 53.3% to 38.9%, the grammaticality of novel test sentences went up by 4% (from 70.7% to 74.7%) when the *gap-link* was added to the EVENT-LINK message.[21] The lower overall sentence accuracy resulted from an increase of clause order related errors. Frequently, the *gap-link* model confused sentence initial NPs (166 times, or 55%) and clause order confusion was also manifest in tense/aspect scrambling in structurally correct sentences. Lacking the *gap-link*, on the other hand, the EVENT-LINK model was using the asymmetric status of the topic/focus element as a cue to clause order. Because the prominence of events was not explicit in the EVENT-LINK message, the model activated both clause initial roles in the WHERE-layer at sentence onset in both the *gap-link* and the *no gap-link* condition. Probing conceptually disconnected thematic role nodes, however, had no overt consequences in the message-lexical system; these nodes were causally inert. If either of the clause initial thematic roles belonged to the gapped element, there was only one NP activated at the WHAT-layer in the *no gap-link* condition, but two in the *gap-link* condition. Thus, there was an unresolvable competition between two NPs with the *gap-link*, whereas the model invariably started out producing the correct clause without the *gap-link*. Once the determiner was produced and fed back to the CWORD-layer, the activated thematic role at the CWHERE-layer always belonged to the set of main clause roles, when there was no *gap-link*. The model could exploit this cue to sequence the next thematic WHERE node in the main clause and thus order events appropriately in production. This explains the higher sentence accuracy in the *no gap-link* condition for the EVENT-LINK message.[22]

The nature of the significant difference in the SIMPLE TOPIC-FOCUS message with and without *gap-link* was more difficult to trace because both models made very few mistakes and there did not appear to be a characteristic error type in the *gap-link* model. In both conditions, the largest class of errors involved confusing double-object and prepositional datives. With *gap-link*, 40% of all errors were of this type, 33% were wrong determiners, the rest was nondescript. Without *gap-link*, 64% of all errors were related to the dative alternation. A crucial difference, however, was that with *gap-link* only 8% dative errors involved the gapped element as indirect object whereas without *gap-link* 100% of all dative errors occurred in the gap position of object-relativized constructions. For instance, a typical dative error of the *gap-link* model was to produce

(30) a. `...that the brother throw -s a mother` instead of
b. `...that the brother throw -s to a mother`

[21] Grammaticality was measured like accuracy but for grammatical categories rather than lexical items.

[22] There were several other construction-specific strategies to use the asymmetric topic/focus status which shall not concern us here.

The model successfully omitted the theme element but produced a double object dative NP instead of a prepositional dative PP recipient. This error points to problems with alternations in general but does not indicate specific problems with the gapped element. The *no gap-link* model, on the other hand, frequently produced

(31) a. `...that a sister give -s a cherry to` instead of
b. `...that a sister give -s a cherry`

Here, the alternation error occurred on the gapped role itself. Furthermore, the *no gap-link* model generated some embedded active/passive errors (12%) which indicated that it was confused about the thematic role of the gapped element. No such errors occurred in the *gap-link* condition. In sum, 76% of all errors in the *no gap-link* condition involved the message focus and were manifest either at the intended gapping site or that of an alternation. This contrasted with only 8% of such errors in the *gap-link* condition. Errors such as (31-a) could have two causes. Either the model had problems distinguishing alternations, or it was uncertain about the thematic role of the gapped element and continued to produce a recipient after the theme (or a combination of both factors). That such errors on the gapped role were absent in the *gap-link* model suggests that uncertainty about the gapped role was the main cause of error in the *no gap-link* model. It also suggests that the *gap-link* helped the model to identify the focus role. Once the topic was produced, the *gap-link* activated the topic and focus roles in the CWHERE-layer. This information was maintained in the CWHERE2-layer memory which accumulated the activation states of previously produced roles. Hence, throughout relative clause production, the HIDDEN-layer was constantly reminded of the identity of the focus role and this information facilitated correct relativization in the *gap-link* condition. Without *gap-link* this cue was not available. This analysis concludes the message comparison and the argumentation in favor of the SIMPLE TOPIC-FOCUS message with *gap-link* as the most suitable meaning representation for the model.

4.5 Discussion

In order to successfully learn and generalize an artificial language with relative clauses, the Dual-path model required message input which encoded

(i) the relative prominence of event participants (alternation bias) in a semantically persistent way,

(ii) the relative prominence of atomic events (theme and comment),

(iii) the topic and focus of a relative clause, endowed with conceptual content (*gap-link*).

Presenting results from this thesis on various occasions, it was suggested to me that the Dual-path model achieved its learning task because strong assumptions were made about the model's meaning representations. Specifically, it was objected that the semantic input was too rich in structure and that this information might not be available to a language learning child.[23]

It is quite clear that a competent speaker must represent the conceptual structure of the intended message for her utterances in similar ways as the Dual-path model. A speaker can be ignorant about the fact that her utterances are difficult to understand or even create ambiguities in comprehension (Ferreira and Dell, 2000), but she must represent fundamental properties of the sentence message herself, such as constructional meaning, topicality, and themehood. In language learning, it could be argued, a child does not have access to such message features when constructing semantic representations. In training the Dual-path model on message-sentence pairs, it was assumed that a child learns a language in situated comprehension in which it can infer many aspects of sentence meaning from a visual environment shared with the speaker through joint attention; aspects such as *who does what to whom*, who is picked out by a relative clause, main clauses convey more important information, etc. In addition to visual information, such as observed actions and events, a child might draw on other kinds of linguistic and non-linguistic information which facilitate the reconstruction of meaning, e.g., prosody, discourse context, and gesture. Given the richness and diversity of information sources available to a child and given children's remarkable capacity to establish reference (Baldwin, 1993) and communicative intentions (Tomasello, 2003) in joint attentional frames, it may be premature to contend that children do not have access to the aspects of sentence meaning which are encoded in the Dual-path model message.

Secondly, although the model received a complete message in training, the semantic information contained in the message is initially not interpreted for the model. Meaning is assigned to message features by the model designer but the model itself must learn to properly interpret these features in the training process. For instance, I referred to feature XX in the event semantics as an agent feature, it was said that active features XX and YY signal an agent and a patient participating in the event, and that the relative prominence of participants was encoded in the message. For the model, however, role features are initially meaningless and indistinguishable from other features (topic/focus, tense/aspect); the message is an uninterpreted linear pattern of activation. The relative prominence of participants was signalled to the model in relation to other constructions (e.g., active/passive), but there was no information in a message in isolation which signalled the order of participants since all participant features were active from the beginning. Alternation parameters as well as role features signalling the number of participants only become meaningful in comparison with other message-sentence pairs. Thus, the model must learn to reconstruct the designer's intentions to encode sentence meaning systematically in the message. This process of reconstruction, or 'making sense', can be interpreted as a way of modelling a child's efforts to construct representations of the conceptual structure of the utterances in its linguistic environment. Furthermore, the

[23] Shimon Edelman in discussion at the *Perception, Cognition and Development* seminar, Cornell University 11/2007.

model received no explicit instructions when and how to use chunks of information in the message. Feedback pertained to mismatches in the predicted sentence form and the model had to infer from this feedback what the message features signalled to use this information appropriately. Because the entire message was given to the model at the beginning of production, it had to learn to use message components selectively in incremental processing. Parts of the message, however, were only becoming fully available to the model once other knowledge had been established. For instance, the model could make use of its role-to-concept bindings only to the extent that it had already learned word meaning from the linguistic input. Moreover, role features in the event semantics did not map one-to-one onto syntactic roles, so the model still had to learn syntactic frames despite being provided with constructional meaning.

And finally, the Dual-path model did not require full access to all message features listed in (i)–(iii) in order to acquire and generalize the grammar of the target language. I trained the model in a condition in which 50% of all message-sentence input pairs were incomplete in that the meaning representation was corrupted. In these pairs, features in the event semantics and role-to-concept bindings in the WHAT-WHERE-system were randomly deleted. Such incomplete messages might adequately reflect developmental stages in which children can only make partial sense of overheard utterances in acquisition.[24] From this defective message-sentence input the model nonetheless learned to produce novel relative clause constructions with 93% accuracy. Although this regime took twice as much training as with complete semantic input, the defective messages did not prevent the model from learning the target language eventually. This suggests that the assumption of complete message input was unnecessarily strong. Partially complete semantic representations would be sufficient for satisfactory learning given that the model could fully 'understand' at least some proportion of its linguistic input, which is not an unreasonable assumption for learning children. In all subsequent experiments I retained complete message input to keep computational time to a minimum.

[24] Developmental stages could also be modelled by incremental training, e.g., starting with 100% defective messages for simple utterances, followed by partially complete simple sentence messages, and so forth. Assumptions made in incremental training, however, might be stronger than the assumption of message completeness for randomized exposure to samples of all sentence structures from the start of learning.

Chapter 5

Model analysis

In this chapter I differentiate between constructions in the input language to the Dual-path model and determine which structures are particularly hard to learn and why. I examine the internal representations the model develops at various layers during learning with the TOPIC-FOCUS message. It will be analyzed whether the model acquires word grammatical classes, phrasal categories, and verb argument structure. I argue that the model constructs sentences incrementally and investigate what the basic planning units are. Furthermore, I take a look at clause-level processing and analyze how the model represents hierarchical sentence structure, attachment and relativization.

5.1 Behavioral analysis

The message comparison of the previous chapter ignored the possibility that model performance might vary across relative clause types. Sentence accuracy was averaged over all complex constructions in the test set. In this section, I compare the learnability of different constructions in terms of the number of syntactic alternations they contain, and in terms of the syntactic role of the topic/focus constituent.

5.1.1 Syntactic alternations

The artificial language of Chapter 4 allowed the expression of transitives in active or passive voice and of transfer events as prepositional or double-object datives. These syntactic alternations express similar propositions but convey different perspectives on the same event. For instance, the passive transitive topicalizes the patient by placing it in sentence initial position. Thus, alternations assign the same semantic but different syntactic roles to participants, and change the order (and therefore relative prominence) of participants in the sentence form. In the model's message, alternations were encoded by an activation-based alternation parameter which placed emphasis on the more prominent participant and biased the model towards choosing one form over its structural alternative. The existence of these alternations in the language creates competition in structural selection for some event types (transitive and dative) but not others (intransitive and oblique). It was hypothesized that this competition of sentence forms for similar messages would lead to greater difficulty in learning constructions which contained syntactic alternations than those which did not.

To test this, multi-clause constructions were partitioned into three classes which together exhaust the input language. In the first class were those complex sentences which contained no alternations in both the main and the subordinate clause. In the second class were sentences which contained alternations either in the main or the subordinate clause. The third class had alternations in both clauses. I regarded the active transitive and the prepositional dative as default structures relative to which the passive transitive and the double-object dative, respectively, alter the order of participants in the sentence form. Hence to 'contain alternations' here means that one or both clauses are passive transitives or double-object datives. For example, the first class comprised complex sentences composed of intransitive, active transitive, oblique and prepositional dative clauses, the third class contained only sentences composed of passive transitive and double object dative clauses.

The Dual-path model is an architectural extension of a simple recurrent network (SRN). These systems are statistical learning models which are sensitive to distributional regularities in the input (Elman, 1991, 1990). In particular, SRN are sensitive to the frequencies of input structures. *Ceteris paribus*, more exposure to structure A than to B leads to better learnability of A. In order to determine whether the Dual-path model had more difficulties with processing alternations we therefore have to rule out a straightforward frequency-based explanation of differential behavior. For example, even if all three classes of sentences occurred with the same frequency in training, there still might be more structural diversity in one class than in another. More diversity entails less exposure to each structure during learning which could lead to decreased accuracy in testing for this class. That there was indeed differential diversity in the three classes of sentences is shown in Table 5.1. The top row indicates the number of combinations of basic constructions in each class. Class three with alternations in both clauses, for instance, contained only sentences composed of passives and double object datives. Hence there were 4 possible combinations of clause types from which these sentences were assembled but 16 for sentences without alternations. Moreover, the number of participants in different constructions can vary (one in intransitives, three in datives) and the more participants, the more ways there are of modifying and relativizing noun

	None	Either clause	Both clauses
Structural combinations	16	16	4
Patterns of relativization	51	58	16

Table 5.1: Structural diversity in the three alternation classes.

phrases. The bottom row of Table 5.1 indicates the number of distinct patterns of relativization in each class. For example, in the first class (without alternations), an active transitive main clause could be combined with an oblique relative clause. Both clauses have two animate participant roles, consequently there were four possible relativization patterns for this clausal combination. Since all of these patterns of relativization must be learned through training, more distinct patterns entail lower individual frequencies. According to Table 5.1 there was more structural diversity in class one than in class two, and more in this class than in class three (in terms of both criteria). Therefore each structure in class one would receive less training than each structure in the other classes, and so forth. To balance this structural diversity the size of the three classes was made proportional to the number of relativization patterns in each class. As usual, the training set consisted of 8.000 simple-clause and 2.000 relative clause sentences.[1] The model was trained as before and tested periodically on 600 novel sentences, 200 from each class, during development. The results of this experiment are shown in Figure 5.1. For improved visibility of the differences between classes, only epochs 30–70.0000 are depicted. At the end of training (after exposure to 100.000 sentences), all classes reached >95% sentence accuracy. Figure 5.1 shows that the learnability of constructions depended on the number of alternations. Sentences with no alternations developed faster than sentences with one alternation, which were learned more easily than sentences with two alternations. Hence, it seems that passive transitives and double

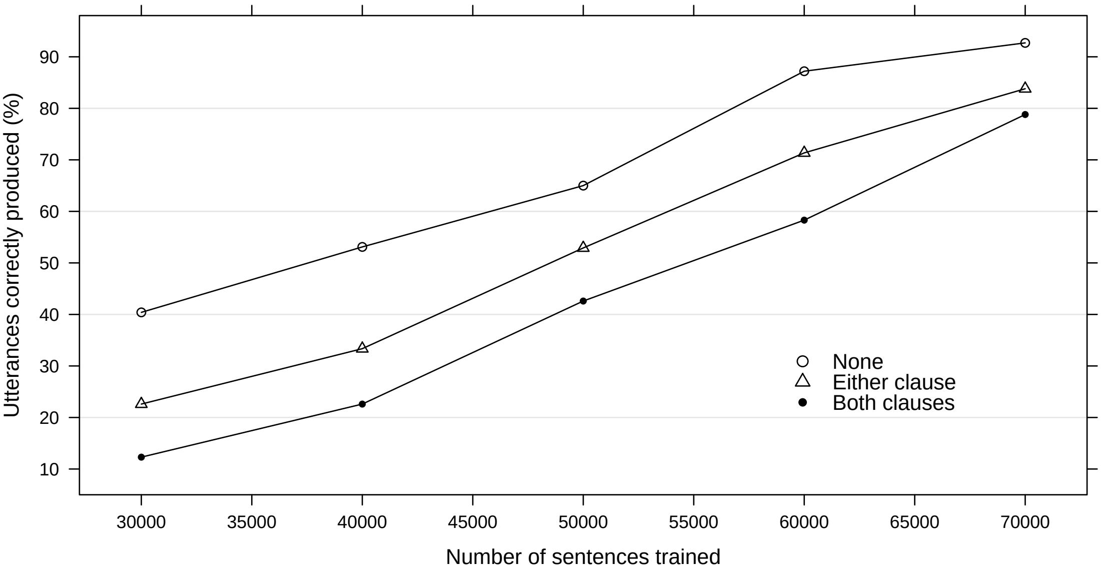


Figure 5.1: Differential learning of complex constructions depending on the number of alternations.

[1] The 2.000 complex sentences contained 816 sentences from class one, 928 sentences from class two, and 256 sentences from class three.

object datives exert a strong influence on the learnability of constructions. The fewer alternations it contained the better a construction was learned.[2]

If structural competition between sentence forms is the reason why alternations were inherently more difficult to learn, this should reciprocally affect active transitives and prepositional datives. To test this, the exact same experiment was conducted for these two structures and also for the combination of passive transitives + prepositional datives and for active transitives + double-object datives. A three-way within subjects ANOVA was conducted at epoch 50.000 with transitive type, dative type and number of alternations as factors and sentence accuracy as the dependent measure. Neither the main effect of transitive type was significant (F(1,9) = 0.21, p = 0.65), nor the main effect of dative type (F(1,9) = 0.98, p = 0.35). There was a significant main effect for the number of alternations (F(2,18) = 157.7, p < 0.001). Moreover, there was a significant interaction of transitive type and number of alternations (F(2,18) = 7.13, p < 0.01), see Figure 5.2. While in sentences with one alternation passive transitives lead to an improvement in accuracy over active transitives, passives were detrimental in sentences with two alternations compared with active transitives.

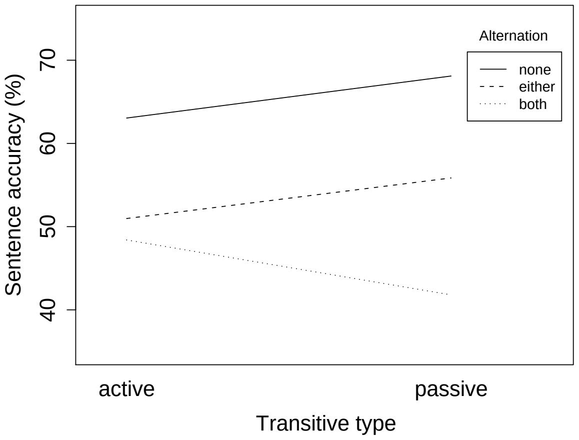


Figure 5.2: Interaction of transitive type and number of alternations.

All four conditions lead to the same ordering of learnability (none > either > both) which suggests that this ordering is robust in that sentences with two alternations are harder to learn, regardless of which combination of transitive and dative structures we consider. This does not entail that the explanation of this ordering is identical for each condition. For analysis, I will return to the condition of Figure 5.1, where only passive transitives and double-object datives occurred in sentences with two alternations.

A frequency-based explanation of the differential behavior was excluded on a constructional level by balancing the training set appropriately. However, frequencies in the training set might be skewed on a sub-constructional level and if the Dual-path model is sensitive to such frequencies, these might generate the ordering of Figure 5.1 in development. For example, the model might be sensitive to frequencies of word category sequences such as THAT VERB DET NOUN and if there are large differences in frequency for such sequences across the three classes of sentences this might influence the learning and processing of alternations. Double object datives, for instance, contain

[2] A further distinction could be made between alternations in the main clause versus alternations in the relative clause. Main clause alternations were learned slightly faster than relative clause alternations.

the sequence VERB DET NOUN DET NOUN which does not occur in any other construction and similarly passive transitives contain the unique sequence VERB PARTICIPLE BY DET NOUN. On the other hand, the oblique construction contains the unique sequence VERB WITH DET NOUN and the intransitive is the only construction with an end-of-sentence marker after the verb form. These two structures occur in combination only in the first class without alternations and did not appear to be particularly difficult to learn.[3] The frequencies of construction-specific substructures can therefore be excluded as an explanation of the model's differential performance.

Bigram statistics

Notwithstanding, it might still be the case that the behavior of Figure 5.1 (page 117) is caused by distributional properties of the input. For instance, the double alternation class might contain bigrams—chunks of two consecutive words or word categories—which are less supported by the training set than bigrams in the no alternation class. To examine this possibility I computed the bigram statistics for the model's training set and determined how well the three tested classes are predicted by these transitional probabilities. First, I converted the training and test sentences into sequences of word categories. A passive sentence such as `the man is chase -par by a dog`, for example, was represented by the sequence DET NOUN AUX VERB PARTICIPLE BY DET NOUN.[4] If s was a test sentence consisting of k word categories, $s = w_1 \dots w_k$, the probability $P(s)$ of s was computed as the product of conditional probabilities between adjacent word categories in s

$$P(s) = \Pi_{i=1}^{k} P(w_i|w_{i-1}) \tag{5.1}$$

where $P(w_i|w_{i-1})$ = (number of times w_i follows w_{i-1})/(number of times w_i occurs) in the training set.[5] Thus, if s is a test sentence, $P(s)$ measures the probability that s is correctly predicted based on the conditional probabilities for bigrams in the training corpus. In other words, the higher $P(s)$ of a novel test sentence s is, the better it is supported by the model's learning environment.[6] Suppose, however, for two sentences s_1 and s_2 we obtain $P(s_1) = P(s_2)$ but s_2 is longer than s_1. Intuitively, in this case s_2 would be better supported because the average bigram probabilities must be higher in s_2 than in s_1. A standard measure which reflects differential sentence length is cross-entropy (Chen and Goodman, 1999). For a set of sentences $S = \{s_1, \dots, s_n\}$, cross-entropy

[3] See the error analysis below.

[4] The prepositions `by`, `to` and `with` were distinguished in tagging because they are characteristic of different constructions.

[5] Since all sentences started with a determiner $P(w_1|w_0)$, the probability that a sentence starts with w_1, was set to 1.

[6] All bigrams in test items occurred in the training corpus, hence no interpolation smoothing was necessary.

$CE(S)$ is defined as

$$CE(S) = \frac{1}{N_S} \sum_{i=1}^{n} - \log_2 P(s_i) \tag{5.2}$$

where $P(s_i)$ is the probability of sentence s_i and N_S is the sum of the lengths of all sentences in S. Cross-entropy is inversely related to the probabilities of definition 5.1. The lower the cross-entropy of a set of sentences, the better this set is supported by the training environment. To see whether this bigram model predicted the order of acquisition in the Dual-path model I calculated the cross-entropy of the three tested classes of sentences (Figure 5.3). The cross-entropy measure predicts that sentences with no alternations should be harder to learn than sentences with one alternation in either clause. These sentences in turn should be more difficult than sentences with two alternations, which had the lowest cross-entropy.[7] This order of difficulty is inverse to the order of acquisition in the Dual-path model. Hence, the cross-entropy model did not explain why the number of alternations correlated with the Dual-path model performance.

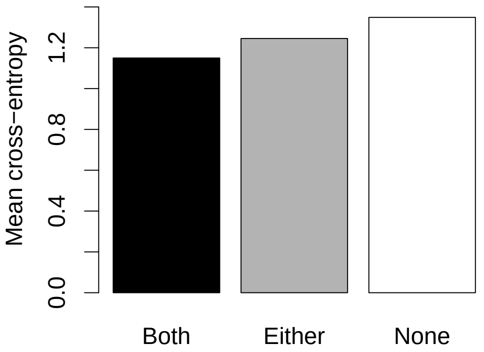


Figure 5.3: Cross-entropy for the three alternation classes.

Frequency of semantic roles

The Dual-path model learns a target grammar from message-sentence pairs and the message is given to the model as input at the beginning of the prediction task. In order to use this semantic input appropriately, the model must learn to activate thematic roles in the intended order in the message-lexical system (see Chang, 2002). This sequencing of roles might be sensitive to the frequencies of chunks of roles in the message-sentence pairs of the training set. Chunks of roles, however, can be construction-specific. For example, the passive transitive is the only construction in the artificial language in which the sentence initial NP is assigned the patient role Y and the object NP the agent role X. To produce a correct passive, the model must first activate the Y role, and later the X role. The transition from Y to X is not supported by any other construction in the target language and this might explain why passive transitives are more difficult to learn. Similarly, the double-object dative includes the subsequences X→Z and Z→Y of roles which are unique to this construction. Other constructions, such as the prepositional dative, require the activation of the X→Y→Z sequence at the WHERE-layer. This sequence

[7] In terms of mean probabilities, the order was $P(both) > P(either) > P(none)$ and $N_{both} > N_{either} > N_{none}$ for sentence length.

is supported by the active transitive X→Y and the oblique Y→Z subsequences. Such semantic differences between constructions might explain why the passive transitive and double-object dative were more difficult to learn and why more of these structures in one sentence are increasingly detrimental. Counting the frequencies of two-role chunks in the training set of the model in Figure 5.1 (page 117), the two tested alternations were the least supported structures in this regard. I balanced two-role chunks to match the support for active transitives and prepositional datives, respectively, across the entire training set. This was achieved by increasing the relative frequencies of simple-clause passives and double-object datives, while leaving all other properties of the training set intact. Training the model in this way resulted in no qualitative and only a small quantitative difference between the two conditions. A two-way within subjects ANOVA was conducted at epoch 50.000 with roles balanced/unbalanced and the number of alternations as factors and sentence accuracy as dependent measure. It indicated no difference in overall performance between the role conditions ($F(1,9) = 1.71$, $p = 0.22$). However, there was an interaction of condition and number ($F(2,18) = 6.77$, $p < 0.01$). There was no difference in mean for two alternations, but an improvement for one alternation and, surprisingly, also for the class of sentences without alternations. Increasing the frequencies of two-role chunks which support the passive transitive and double-object dative did not lead to improved performance for double alternations and even increased the differences between the three classes of sentences. This suggests that the distribution of chunks of semantic information does not explain the differential behavior of Figure 5.1.

Error analysis

In the previous sections I excluded a number of frequency-based explanations for why alternations are particularly hard to process by the Dual-path model. Constructional frequencies and relativization patterns were equated in training. Neither substructure frequencies or bigram statistics in the corpus, nor the frequencies of two-role chunks in the message explained the model's behavior. These negative results indicate that alternations are inherently harder to process because similar meaning representations are mapped to different sentence forms. To assess this possibility, the model's error profile for each class of sentences was examined. Structural competition for similar messages in alternations should be manifest in higher error rates for transitive and dative test items and in the amount of structural conversion between alternations in the model's output, e.g., the conversion of active transitive target sentences to passive transitives. For each class, the first 50 errors were inspected and classified with respect to the basic construction type in which they occurred (e.g., intransitive versus dative), in which clause they occurred (main or relative clause), and whether they were conversion errors or not.[8] A typical conversion from passive to active (labelled P/A error), for instance, occurred when the model produced the subordinate verb after the pronoun in a subject-relativized clause, i.e.,

[8] All examined errors occurred after the model had experienced 80.000 training sentences. A late epoch was chosen for analysis because errors were more distinctive.

(1) a. `the mother is push -ing a girl that kick ...` instead of
b. `the mother is push -ing a girl that a dog is kick -par by .`

The error profiles for each class of sentences are shown in Table 5.2. The 'construction' column indicates the clause type and location of the sequentially first error in incorrect productions. The 'conversions' column indicates the percentage of X-type errors which was characteristic of a conversion to structure Y for each X/Y error listed. For instance, in the no alternation class a total of 18 errors occurred in active transitive clauses out of which 83% were conversion errors to passive transitives. The remaining 17% of errors were either lexical or nondescript. In this class passives and double-object datives were not tested and 84% of all errors occurred in active transitives and prepositional datives. 81% of these transitive and dative errors were conversions to the corresponding alternation structure. In the class with one alternation in either clause, the majority of errors occurred in datives and the P/A and DO/PD conversion rates were highest. In the class with two alternations, the majority of errors were conversions to the structures competing with the construction types in the tested sentences.

No Alternations

	Construction						Conversions			
Clause	I	A	P	O	PD	DO	A/P	P/A	PD/DO	DO/PD
Main	0	7	0	2	7	0	5	0	5	0
RC	2	11	0	4	17	0	10	0	14	0
Percent	4%	36%	-	12%	48%	-	83%	-	79%	-

One Alternation

	Construction						Conversions			
Clause	I	A	P	O	PD	DO	A/P	P/A	PD/DO	DO/PD
Main	0	3	2	1	7	6	0	1	1	4
RC	0	4	6	3	10	8	1	4	7	6
Percent	-	14%	16%	8%	34%	28%	14%	63%	47%	71%

Two Alternations

	Construction						Conversions			
Clause	I	A	P	O	PD	DO	A/P	P/A	PD/DO	DO/PD
Main	0	0	5	0	0	11	0	3	0	5
RC	0	0	18	0	0	16	0	10	0	12
Percent	-	-	46%	-	-	54%	-	57%	-	63%

Table 5.2: Error profile for the three alternation classes. Labels mean: I = intransitive, A = active transitive, P = passive transitive, O = oblique, PD = prepositional dative, DO = double-object dative, RC = relative clause.

These error distributions suggest that the model had difficulties discerning the transitive and dative pairs of alternations which lead to more errors in these constructions and ultimately to an increasing amount of errors in those classes which contained more alternations (Figure 5.1, page 117). The difficulty with alternations is rooted in the similarity of these structures in the message input to the model. Meaning representations for alternations were composed of the same semantic features in the event semantics and the same role-to-concept bindings in the WHAT-WHERE-system. They only differed in terms of the relative strength in the activation of participant features. This alternation parameter biased the model to select the intended transitive or dative structure but the bias was not strong enough to prevent conversion errors from occurring. Alternations complicate the meaning-to-form mapping the model has to acquire since they invert the order of semantic roles in the sentence form. To produce a correct passive, for instance, the model must begin production by sequencing the patient role. Once the patient has been produced, this constituents' thematic role is maintained in the memory of the CWHERE-system due to feedback, and this facilitates the sequencing of the agent role later on. Hence, semantic information is accreting during production and guides the model in structural selection. At the onset of a clause, however, the model can rely solely on its message input. Thus early uncertainty caused by highly similar messages leads to clause-initial errors as in sentence (1-a) above. At the error position the model knows that `girl` is the focus and the agent of the relative clause but is confused about its grammatical role, leading it to produce `that kick...` (`girl` = agent, subject) rather than `that a dog kick...` (`girl` = patient, object), or the target sequence `that a dog is kick...` (`girl` = agent, object).

To summarize, distributional properties of the input did not explain why sentences with more passive transitives and double-object datives were more difficult to learn. In transitives and datives the same role features can get mapped to distinct grammatical roles and this difference was encoded in the conceptual structure of the message by signalling the relative prominence of participants. Alternation messages were therefore very similar but mapped to dissimilar sentence forms. This complication made alternations inherently difficult to learn which was traceable in the high percentage of structural conversions in the model's error profile and this explains the differential performance for alternation classes witnessed in Figure 5.1.

5.1.2 Topic and focus

In the literature on language acquisition, relative clauses have received wide attention. A large number of studies have investigated the order of acquisition in English-speaking children, measured mainly in comprehension. In these studies it is customary to characterize relative clause types in terms of the grammatical role of the topic and focus elements. Four classes of relative clause structures can be distinguished in this way:

- *SS-relatives* in which the main clause subject is the topic of the relative clause and the relative clause subject the focus.

- *SO-relatives* in which the main clause subject is the topic of the relative clause and the relative clause object the focus.

- *OS-relatives* in which the main clause object is the topic of the relative clause and the relative clause subject the focus.
- *OO-relatives* in which the main clause object is the topic of the relative clause and the relative clause object the focus.

An example of each structure from the artificial language is given below.

(SS) `the girl that throw -ed the orange to a cat fall -ed .`

(SO) `the woman that a boy teach -s was push -par by a dog .`

(OS) `a mother was give -ing a cookie to a nurse that run -s with a man .`

(OO) `the brother kick -ed the dog that the teacher is approach -ing .`

To test children's comprehension abilities, most studies employed either a sentence repetition task in which children had to repeat sentences with relative clauses from the experimenter, or an act-out task in which children heard such sentences and were asked to act out their meaning with toys. Based on children's error rates the order of acquisition between SS-, SO-, OS-, and OO-relatives could be estimated. Table 5.3 shows the results from four studies which have received the most interest and a fifth more recent study. Obviously, there is large variation in these findings, e.g., OS-relatives are the easiest or the hardest structures, depending on the particular study. Despite plenty of psycholinguistic experiments on this issue, no clear picture has evolved to date. In similar vein, many interpretation strategies, which perhaps are used by children in relative clause processing, were proposed as an explanation of differential behavior:[9]

Study	Result
de Villiers, Flusberg, Hakuta, and Cohen (1979)	(OS, SS) > OO > SO
Sheldon (1974)	(SS, OO) > (SO, OS)
Smith (1974)	OS > SS > OO > SO
Tavakolian (1981)	SS > (OO, SO) > OS
Kidd and Bavin (2002)	(OS, OO) > SS > SO

Table 5.3: Children's comprehension of relative clauses. X>Y indicates that X was easier to comprehend than Y.

(i) *Non-interruption hypothesis* (Slobin, 1973): relative clauses interrupting the main clause are more difficult (prediction: (OS, OO) > (SO, SS)).

(ii) *Conjoined-clause hypothesis* (Tavakolian, 1981): children interpret sentences with relative clauses as a conjunction of two simple sentences (prediction: SS > (OO, SO) > OS).

[9] Cf. Diessel (2004) and the detailed discussion therein.

(iii) *Parallel-function hypothesis* (Sheldon, 1974): children assign the same syntactic roles to the topic and focus of a relative clause (prediction: (SS, OO) > (SO, OS)).

(iv) *NVN-schema hypothesis* (Bever, 1970): children have less difficulty with relative clause constructions which follow the noun-verb-noun pattern of simple transitive clauses (prediction: (SS, OS) > (SO, OO)).

(v) *Filler-gap hypothesis* (Wanner and Maratsos, 1978): processing difficulty varies with the distance between the topic and focus of the relative clause (prediction: (SS, OS) > (SO, OO)).

In order to determine where in this complicated landscape of results and explanations the Dual-path model fits in, the model was trained as usual with a set of 10.000 sentences, 20% of which contained all of the four types of relative clauses (SS, OS, SO, OO). As in the alternation experiment the input frequencies were balanced so that the number of training items divided by the number of distinct constructions in each class was equal. The model was periodically tested on a set of novel sentences containing 200 items of each type, drawn uniformly from the language. Figure 5.4 depicts the results from this comparison (averaged over ten model subjects).[10] In this condition, I obtained the processing hierarchy SS > (SO, OS) > OO. Strictly speaking, this ordering is not in line with any experimental data or any single explanatory hypothesis mentioned above.

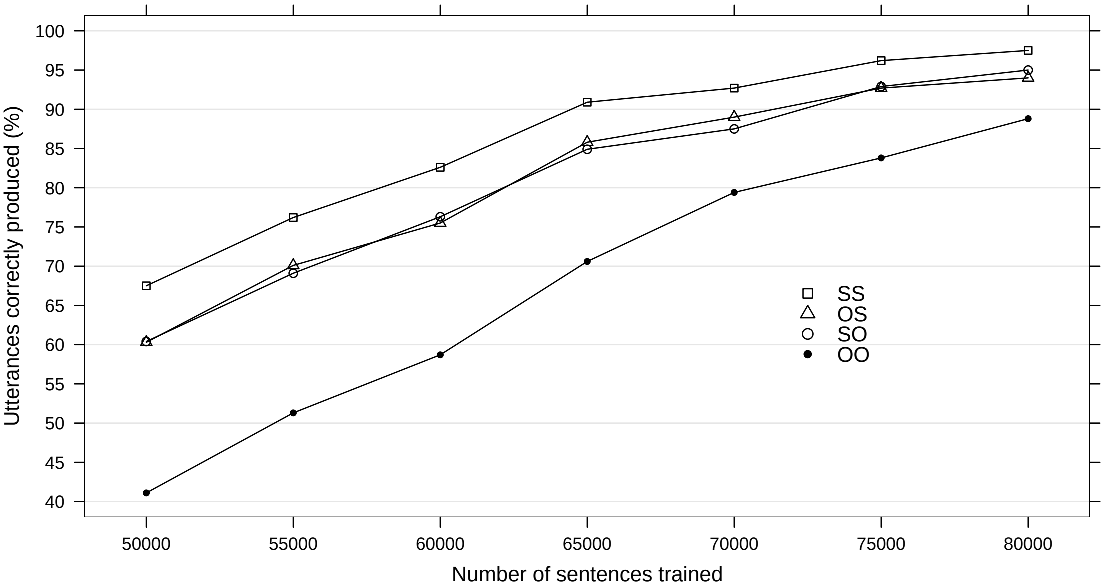


Figure 5.4: Comparison of SS-, SO-, OS-, and OO-relative learning.

[10] The model was trained for 100.000 epochs and all constructions eventually reached ceiling (>95% sentence accuracy). For better visibility of the contrasts, only epochs 50–80.000 are displayed.

The large diversity in these data, however, indicates that most likely there is not a single factor responsible for children's differential comprehension but rather some complex interaction of the factors (i)–(v) and possibly several others. Overall, relative clauses attached to the main clause subject (SS, SO) were learned faster by the Dual-path model than relative clauses attached to the main clause object (OS, OO). This behavior appears to be inconsistent with the non-interruption hypothesis which predicts that object attachment should be easier because the main clause is uninterrupted by intervening material. Notice, however, that object attachment does not guarantee an uninterrupted main clause. In my language, ditransitive and prepositional datives could have both objects modified and relativized. As a consequence, the main clause could be disrupted by a relative clause, but still classify as an OO-relative:

(2) `a man bring -s a cat [that a woman give -ed a toy to] the apple .`

Given the results from the previous section on alternations we know that such structures are particularly hard for the Dual-path model and the interruption of the main clause might contribute to this difficulty. Because the non-interruption hypothesis does not predict the relative difficulty of such sentence types it might be premature to conclude that the low performance on OO-relatives in the model is at odds with this hypothesis. The model's behavior is partially consistent with the conjoined-clause hypothesis in that SS-relatives were the fastest structures to develop in the model. It is also partially consistent with the parallel-function hypothesis which predicts that SO- and OS-relatives should cause the same amount of difficulty because topic and focus assume different syntactic roles, and that both are harder than SS-relatives which assign the same role to the head and relativized element. Furthermore, the performance in Figure 5.4 is partially consistent with the NVN-schema hypothesis in that SS-relatives are learned faster than SO-relatives and OS-relatives are learned faster than OO-relatives. In the same regard, the model's behavior is partially consistent with the filler-gap hypothesis.

So where does this partial concordance on all fronts leave us? It would be desirable to determine the influence of each of the proposed hypotheses (i)–(v) in relative clause processing within the framework of the Dual-path model. A number of factors, however, make this a very complicated endeavor not to be undertaken here. First of all, the experimental studies from which these hypotheses are derived were studies of relative clause comprehension, not production. Unlike comprehension, production does not require interpretation strategies and online integration of overheard linguistic material. It is therefore doubtful whether the Dual-path production model can be utilized to assess, e.g., the conjoined-clause or the filler-gap hypothesis in a straightforward way. Secondly, the lack of reliable, replicable data poses serious methodological problems for a computational modelling approach. One explanatory route would be to make the model match the behavioral data and work backwards from there to identify the processing factors which bring about this behavior. But since there is no consensus on the order of relative clause acquisition the model cannot be calibrated to match the data in the first place. In a bottom-up approach the model could be built on realistic frequencies of relative clause types in child-directed speech. This would yield testable predictions and model-based explanatory strategies which might help evaluate the significance of the factors (i)–(v). Such frequency data is difficult to obtain but in light of the large variation in the comprehension data, a corpus-based modelling account might be more promising than a top-down approach. Third, this variation in the child data could be an indication that the topic/focus-taxonomy of SS-, SO-, OS-, and OO-relatives might be too coarse. Other factors than the subject/object distinction have been shown to strongly influence relative clause processing:

(a) the grammatical type of modified and relativized objects (de Villiers et al., 1979; Keenan and Hawkins, 1987; Diessel and Tomasello, 2005)

(b) the animacy of head nouns and relative clause subjects/objects (Traxler et al., 2002; Mak et al., 2002; Kidd et al., 2007; Gennari and MacDonald, 2008)

(c) semantic determinacy, verb class and relative clause voice (Gennari and MacDonald, 2008)

(d) pronominal relative clause subjects (Gordon et al., 2001; Warren and Gibson, 2002; Kidd et al., 2007)

(e) the distributional properties of relative clause types and the frequencies of substructures (Reali and Christiansen, 2007a,b)

(f) long-term linguistic experience and familiarity with different relative clause types (MacDonald and Christiansen, 2002; Wells et al., 2008)

This range of findings suggests that the classification of relative clauses in terms of the grammatical role of the topic and focus may not be fine-grained enough and that a universal hierarchy for SS-, SO-, OS- and OO-relative processing and development might not be obtainable.

Another source of difficulty for assessing hypotheses (i)–(v) lies in the complexity of the model itself. A large number of factors might influence differential learning in the model. For instance, it was argued in the previous section that alternations are inherently difficult to process, so the number of alternations in each class (SS, SO, OS, OO) will have an effect on the ordering in testing. In the experiment of Figure 5.4, the number of alternations in training predicted the hierarchy SS > SO > OS > OO in testing. Because the Dual-path model is a statistical learning mechanism it is sensitive to distributional properties of its input on many different levels. Constructional diversity in each class was balanced in the experiment, but substructure frequencies and in particular bigram statistics might influence learning and development as well. Cross-entropy in the training set, for example, predicted the order OS > SS > OO > SO. Furthermore, the model's performance might be dependent on the total amount of training for each semantic role and the frequency of each role in the constructions of each class. This relation predicted the order SS > SO > OS > OO in the above experiment. SS-relatives tend to be shorter than OO-relatives because they can be composed of intransitives, and sentence length is perhaps also an important factor because shorter sequences put less strain on the model's working memory. Sentence length predicted the order SS > (SO, OS) > OO, which was actually observed (Figure 5.4).

Test items in the four relative clause classes could be composed of all possible combinations of basic constructions, e.g., an SS-relative could have an intransitive main clause and a dative embedding, or a passive transitive main clause and an active transitive embedding. The model's performance varied for different structures within a single class. The studies of Table 5.3 also differed in the kinds of structures that were tested in comprehension which might be another reason for the variance in these results. When the model was tested on sentences which only contained oblique main clauses and active transitive relative clauses I obtained the processing order (SO, SS) > OS > OO.[11] Moreover, the model's performance might depend on the kinds of basic constructions in the artificial language itself. For instance, there might be subtle effects of interference and similarity between distinct complex structures, and between these and simple-clause constructions. Contrasts in the SS > (SO, OS) > OO processing order might be due to such patterns of interference and facilitation in the input language, and not due to the inherent difficulty of any one class of constructions in isolation.

The variety of factors which might have an influence on learning and processing makes it difficult to trace the cause of the model's differential performance on SS-, OS-, SO- and OO-relatives. It also makes it difficult to isolate and test the factors proposed in hypotheses (i)–(v). In later chapters I will take up this issue again and analyze two aspects of relative clause development which are related to the processing hierarchy of this section. Chapter 6 will examine the model's differential behavior on center-embedded and right-branching structures. In Chapter 8, I investigate how the grammatical role of the relativized element, and in particular the type of object (direct, indirect, oblique), influences the order of relative clause acquisition in the model.

5.2 Representational analysis

So far I have only looked at the overt linguistic behavior of the model by measuring its performance for distinct constructions in various conditions. I now want to gain a better understanding of the specific function each of the two model pathways—the message-lexical system and the sequencing system—performs in order to generate this behavior. For this purpose it is helpful to analyze the internal representations which have developed in each pathway during learning. In the message-lexical system, I will look at the states of the WHERE-layer in the course of sentence production. At this layer, the model has to sequence thematic roles in order to activate sentence-specific content in the WHAT-layer. Inspecting the activation states of the WHERE-layer during production, we can track incremental structural selection and obtain some insights into the model's units of planning.

[11] The test items in the study of Kidd et al. (2007) were of this form.

In the sequencing system I examine the representations at the COMPRESS-layer which immediately precedes the WORD-output layer. The question was whether this layer acquired knowledge of word categories which could be interpreted as syntactic representations used by the sequencing system in sentence processing.

Both pathways of the model are fed by the central HIDDEN-layer. I investigate whether this layer represented traditional phrasal categories and the argument structure of basic constructions. Finally, I trace the representational similarity of clauses in different sentential structures at the HIDDEN-layer.

5.2.1 Lexical categories

The COMPRESS-layer consisted of 20 units and formed an information bottleneck because it was considerably smaller in size than the HIDDEN- and WORD-layers. All activation in the sequencing pathway which propagates to the WORD output must pass through this layer. Consequently, the COMPRESS-layer is forced to develop generalizations which are independent of specific lexical items. Activation at the COMPRESS-layer was recorded for a single trained model (epoch 100.000) while producing a set of 4.000 test sentences, half of which contained relative clauses. The activation vectors where averaged by word category and verb class and quantized into five different activation levels. Black squares indicate an average unit activation of 0.5 or more, white squares indicate activation of 0.1 or less, and different grey-scale squares lie in between (table 5.4, page 130).[12] Because unit activation was averaged, dark squares convey more useful information than lighter squares. Black and dark squares can be found in almost every column, hence there was little redundancy in the COMPRESS-layer. Representations for each row are distributed over several units which means that no single unit encoded one category and different categories shared individual units. Nouns almost exclusively used units C6, C8, and C15. Verbs mainly used units C1, C6, C11–C14, and C16. Units C6 and C12 were shared by all verb classes, whereas other units strongly distinguished between verb classes. For instance, units C13 and C16 separate intransitive and oblique verbs from transitive and dative verbs. Activation patterns for intransitive and oblique verbs appear to be very similar, only units C8 and C11 weakly distinguish both classes. This is because several verbs in the artificial language could occur in intransitive as well as oblique constructions, e.g., `run` and `jump`. Similarly, transitive and dative verbs were only distinguished by unit C14. Although there was no overlap between these verb classes, some transitive verbs used in the artificial language could also occur in prepositional dative frames in natural language, e.g., `kick` in `the man kicked the ball to the boy`. Moreover, the active transitive and prepositional dative construction shared the same initial sequence of role features (XX=agent and YY=patient/theme) in the event semantics. Hence, it was appropriate for the model to represent these two verb classes in similar ways at the COMPRESS-layer.

Unlike nouns and most verbs, functional constituents such as determiners and prepo-

[12] This analysis follows the procedure proposed in Chang (2002).

Syntactic categories and verb classes	Compress layer units																			
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
Determiners	▒	■	■	■			▓		■	■	░		▓	▓		░		░		░
Nouns	░					■		■							■			░		
Prepositions			▒	▓	░	▓	■	■	■	▓	▓			▓	▒	▓	▓	▓		
Auxiliaries	░	▒	▓				░		▒	░	■	░	▓	■		▓		░	▒	▓
Relative pronoun			■	▒							░			■		░	▓			
Intransitive verbs		░	░		░	▓		▓	░	░	■	■		▓		■		▒		▓
Transitive verbs	▓				░	■						■	■	▒					▓	
Dative verbs	■	░	░			▓			░			■	■	■					▒	
Oblique verbs			░		░	■			░		▓	■		▓		■		░		░
End of sentence		▒	░	▒		░	■	■		▒	■					■				■

Table 5.4: Mean activation at the COMPRESS-layer by syntactic category and verb class.

sitions recruited more resources by activating a larger number of units. This is an indication that the model mainly relied on the sequencing system to produce these constituents (cf. Chang, 2002). An exception is the pronoun `that` which effectively activated only two units, C3 and C14. The production of this constituent must be guided by the sequencing system exclusively because it is not part of the sentence message in the WHAT-WHERE system. Thus, function and content words were processed in different pathways in the model. Evidence for this separation comes from a number of studies in psycholinguistics, neuroscience, and aphasia research which have shown that there is a double dissociation between these word classes in human syntactic processing (Goodglass and Kaplan, 1983; Pulvermüller, 1995; Osterhout, 1997, Brown et al., 1999). In contrast to other functional elements, however, the *topic* feature in the event semantics constrained the position of the pronoun within a sentence. In general, the positional variation of `that` in complex sentences was higher than that of prepositions but the semantic information in the message was more explicit for the pronoun position. This allowed the sequencing system to invest fewer encoding units at the COMPRESS-layer. The average activation of the two pronoun units was 0.92 (C3) and 0.99 (C14), respectively. Thus neither unit represented the difference between subject-modifying and object-modifying relative clauses. To probe this difference, the activation states of the COMPRESS-layer were recorded separately for 1.000 subject-modifying and 1.000 object-modifying relative clauses. The results are depicted in Table 5.5, along with the activation pattern averaged over the entire multi-clause fragment from Table 5.4. While the pronoun in

Complementizer	COMPRESS-layer units									
by structure	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Subject-modified			■							
Object-modified			■	▩						
Averaged complex			■	▨						
Complementizer	COMPRESS-layer units									
by structure	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
Subject-modified	░		░	■			░			
Object-modified	░			■		▩	■	░		░
Averaged complex	░			■		░	▩			

Table 5.5: Differential activation of `that` for center-embedded and right-branching structures at the COMPRESS-layer.

subject-modifying relative clauses used only units C3 and C14, in object-modifying relative clauses it additionally used C4, C16 and unit C17 in particular. This unit was highly active (0.73) when `that` initiated a relative clause which attached to a sentence object and almost silent (0.10) when the relative clause attached to a sentence subject. In this manner, unit C17 encoded crucial structural properties of hierarchically distinct sentential constructions. The structural decision between subject versus object attachment that the model had to make in sentence production was manifest in different activation patterns at the pronoun position. Consequently, the COMPRESS-layer was not limited to representing local within-clause syntactic information such as word category and verb class, but also represented structural distinctions between sentence types.

Overall, the activation patterns for nouns, verbs and functional constituents reported here for multi-clause utterances conformed to those found in Chang (2002) for single-clause utterances. Distributed representations developed at the COMPRESS-layer and encoded syntactic categories and verb class information in essentially the same way in both analyses. In this respect, the model scaled well for increased structural complexity in the input language. This can be interpreted as evidence for the robustness of the syntactic representations developed by the Dual-path model in both studies, and as a justification for the architectural assumption of separate processing pathways.

5.2.2 Thematic role sequencing

During sentence production, the model activates chains of word categories in its sequencing system. For example, when producing a grammatical subject-modified, active transitive main clause and an oblique object-relativized embedding we would observe a chain of activation corresponding to the sequence of categories DET NOUN PRON DET NOUN OVERB PREP TVERB DET NOUN at the COMPRESS-layer.[13] In the message-lexical system, on the other hand, the model assigns conceptual content to the positions in these word category sequences. This is achieved by activating a corresponding chain of thematic role units in the WHERE-layer. These units then propagate activation along the role-to-concept bindings to activate conceptual units in the WHAT-layer. Units in this layer represent the lexical-semantics of words. They project to the WORD-output layer and activate a word form. At the output layer, both pathways are joined and they compete for each sentence position. While the message-lexical pathway activates possible word continuations, the sequencing pathways activates possible word category continuations. Jointly, both pathways predict the next word in a production sequence. For instance, the message-lexical system might be in a state of uncertainty and activate several words from different lexical categories. If the sequencing system has already learned the syntactic frame for the target construction it will activate the correct word category for the current position. The combined activation from both pathways will then support the selection of the appropriate word at the output layer. Words which are erroneously activated by the message-lexical system are not supported by the sequencing system and loose the competition for production.

In order for this process to work, the message-lexical system must learn to sequence thematic roles at the WHERE-layer in the right order. To understand the details of this process, the activation states of the WHERE-layer were recorded for 100 sentences with

[13] DET = determiner, PRON = pronoun, OVERB = oblique verb, PREP = preposition, TVERB = transitive verb.

an active transitive main clause and a subject-modifying, object-relativized oblique embedding such as

(3) `a woman that the boy play -ed with is hit -ing the father .`

These activation states were averaged over ten model subjects at epoch 100.000 and plotted in Table 5.6. Thus a representative thematic role sequence was obtained for such structures. At the beginning of the sentence, the model activated the agent node 0X for the transitive clause-initial NP. If the oblique clause had been the main clause instead, the patient node 1Y would have to be activated first. Hence, the slight activation of the 1Y node indicates that the model was not entirely certain about the order of clauses within the sentence when producing the sentence-initial determiner.[14] This uncertainty vanished, however, when the subsequent noun was produced since the model assigned a unique role to it (0X again). At the pronoun position, there is sporadic activation across the WHERE-layer. The pronoun carries no thematic role, hence no WHERE-layer unit is fully active (black square) when `that` is produced. The sequencing system alone is responsible for filling this slot with a lexical item. At this crucial point in structural selection the message-lexical system slightly activates the action role 0A and the patient role 0Y. Both choices make sense for a transitive clause which is not disrupted by a relative clause. In this case the action role 0A would have to be sequenced now,

Word	WHERE layer (thematic roles)							
Category	0A	0X	0Y	0Z	1A	1X	1Y	1Z
DET		■					░	
NOUN		■						
PRON	▒	▒	░				▒	
DET					▓		■	
NOUN						░	■	
OVERB					■	▓	░	
PREP	▓				▒	▒		░
TVERB	■		░			▒		
DET	░		■					
NOUN			■					

Table 5.6: WHERE-layer activation states for 100 sentences with transitive main clause and oblique relative clause, averaged over ten model subjects at the end of training.

[14] Recall that the WHERE-layer units indexed by 0/1 were not dedicated to clause order, i.e., there was no default order. Although Table 5.6 depicts mean activation over 0-1 ordered sentences only, the reverse order 1-0 was also admissible in training and testing.

and given that the pronoun slot could have been the action position already, the patient role 0Y would be next. The model is also re-activating the transitive agent role 0X of the constituent which is modified by the relative clause and it is projecting the relative clause-initial oblique patient role 1Y. In other words, the embedded clause subject is already available at the pronoun position and this is in line with the optional use of `that` in English relative clause constructions. The non-specific activation vector observed at the pronoun position indicates that the WHERE-layer is in a state of uncertainty in which all possible thematic role continuations (given the construction message in the event semantics) are moderately active. To put it differently, multiple available thematic roles compete for this position in the message-lexical system. The competition for content words among various roles is overwritten by the structural choice made in the sequencing system which leads to the production of the functional pronoun. This behavior is consistent with the lexical-syntactic interaction model proposed by Ferreira and Dell (2000) to account for various experimental results on `that` omission in complement structures. Moreover, it shows that the model has retained syntactic flexibility despite its complete message input. It has not planned the entire sentence structure ahead of production but incrementally assigns thematic roles in the time course of producing an utterance (cf. Chang, 2002). Grammatical encoding is not complete before production begins, but sentences are constructed in a piecemeal fashion from start to finish. The sequential activation and deactivation of thematic roles at the WHERE-layer can be characterized as an attentional spotlight which selectively moves over event participants signalled by the message. This incremental behavior is in line with a study by Griffin and Bock (2000) who found concurrence between eye-movement and structural choices in production.

Once the pronoun is produced, all main clause units are switched off and the attentional focus shifts to the embedded clause roles. The incremental nature of processing in the Dual-path model can again be witnessed here in that both the action role 1A as well as the relative clause subject 1Y are activated in the determiner position. Both a subject- as well as an object-relativized oblique embedding are options for the model at this point. Activation of 1A corresponds to a subject-relativized structure and activation of 1Y to an object-relativized structure. Since the oblique subject role 1Y is slightly more active, the corresponding NP gets produced and thus the intended object-relativized oblique clause is initiated by a determiner. The model's choice solidifies at the next step in processing. The determiner is fed back to the model's input layer and activates the 1Y role in the CWHERE-system which signals to the model that an object-relativized clause was intended. Consequently, the action role 1A is deactivated, the 1Y role is fully activated and the subject noun is sequenced at the output layer. The slight activation of the embedded unit 1X can be viewed as a semantic bias towards associating sentential subjects with an agent role.

In the verb position, the embedded action role 1A wins the competition. There is residual activation of the 1Y role and rather strong activation of the 1X role. The 1X role is not linked to conceptual content in oblique constructions so there are no causal consequences to activating this role. Most likely, the model is preparing to sequence this role in the post-verbal slot because in the passive transitive construction the agent role 1X succeeds the 1Y role. If this is the correct explanation, this behavior indicates that the Dual-path model is sensitive to statistical regularities in two-role chunks within clauses (cf. Subsection 5.1.1 above).

In the position of the preposition `with`, the main clause action role 0A is most active. For the majority of subject-modifying, object-relativized sentences in the training corpus this would be the appropriate continuation after the embedded verb. There is also residual activation of the embedded action role 1A and the agent role 1X. Yet, again the sequencing system enforces the production of the preposition here. Least active is the oblique object role 1Z in the current sentence position. Since the thematic role of the gapped element is linked to conceptual content in the WHAT-WHERE-system, the model had to suppress activating this role. The preposition completes the relative clause and attention shifts back to the incomplete main clause. This re-entry causes some difficulties, comparable to the clause-order uncertainty at the beginning of the sentence, as the transitive verb is the only content word in the sentence for which active roles can be found in both clauses. Finally, the correct patient role 0Y gets unambiguously assigned to the clause-final direct object NP which suggests that queueing the production of main clause participant roles over the disrupting relative clause is not a source of difficulty for the model.

By analyzing activation patterns at the WHERE-layer for the sequence DET NOUN PRON DET NOUN OVERB PREP TVERB DET NOUN of word categories we could identify four interrelated characteristics of the Dual-path production model. Thematic role assignment is driven by activation-based *competition* among simultaneously active roles. In most sentence positions, multiple roles compete for the next slot by showing some activity and the most active role wins. The active range of possible roles in each sentence position reflects the model's experience of structural types in the language. For instance, the activation of the agent role 1X after sequencing the 1Y role inside the embedding shows that the model considers a passive transitive as a possible relative clause, despite having received an input message for an oblique clause. The strength of activation of competing roles which are not selected is correlated with the training frequencies of alternative structures. Secondly, the model assigns thematic roles *incrementally* which indicates that structural choices are being made in the course of production. Incremental processing is visible at various points in the test structure and most pronounced at the relative clause-initial determiner. Here the model has to choose between a subject-relativized and an object-relativized embedding. Both types of relative clauses are available and considered plausible by the model in that the embedded action role and the role of the subject NP are highly active. The tentative activation and subsequent deactivation of roles shows that structural selection follows a step-by-step regime and is not fully planned in advance. When multiple roles are equally active at function word positions, in particular the pronoun, the sequencing system imposes syntactic constraints and regulates structural choice. Thus the message-lexical and the sequencing pathway *interact* in the course of generating a sentence. As a consequence of activation-based competition, incremental processing, and pathway interaction, the Dual-path model remains syntactically *flexible* throughout sentence production and is not rigidly committed to a preconceived grammatical encoding of its semantic input.

Hierarchical planning

A long-standing controversy in language processing concerns the question what the basic units of sentence planning are. On the serial order account, planning is based solely on the transitional probabilities between units of the same type (e.g., words). Co-occurrence frequencies in the experience of speakers determine the strength of sequential connections between these units. On the hierarchical account, the syntactic structure of a sentence is planned over larger, hierarchically connected units such as finite clauses. Evidence for the latter account comes from a number of sentence production studies (Boomer, 1965; Ford and Holmes, 1978; Holmes, 1988; Garrett, 1988; Bock and Cutting, 1992). These studies differed in their methodologies but a common theme was to look at error positions and frequencies in the production of structurally complex sentences in various experimental conditions. Ford and Holmes (1978), for instance, asked subjects to respond to tones while they spoke and measured the reaction times (RT) across sentence positions. They found that RTs were longer at the end of clauses than at the beginning. Longer reaction times indicated positions of increased processing load at which speakers planned ahead of the current speech and they interpreted these findings as evidence for clauses as a crucial planning unit. Holmes (1988) found more pauses and hesitations at the beginning of a sentence and before embedded clauses in spontaneous speech compared with reading out loud, indicating the clausal planning positions in speech. Bock and Cutting (1992) elicited more agreement errors in sentences in which the head noun-verb dependency was interrupted by a prepositional phrase rather than an embedded clause. They argued that this error profile supports a hierarchical account because phrasal material in the same clause causes more interference between active constituents than material which belongs to a different clause. The processing focus shifts to the embedded clause which then receives priority over material outside this clause, suggesting that clauses are basic planning units.

These findings, and hierarchical planning in general, may at first appear to be inconsistent with production in the Dual-path model which generates utterances on a word-by-word basis and makes decisions about thematic role assignment incrementally. Unlike other sequential learning models, however, the Dual-path model is not relying exclusively on transitional probabilities in word-to-word prediction. The model also receives semantic information in the form of message input before production begins. Since this message is provided non-incrementally it can in principle form the basis of hierarchical planning.

Competition among active thematic roles in the WHERE-layer indicates junctures of uncertainty and more competition increases the likelihood of production errors. Thus, we can interpret sentence positions at which there is strong competition as hesitations or loci of planning in the model. In Figure 5.6 (page 133) there is more competition in the relative clause than in the main clause. In a radically incremental, serial order model we would not expect such an asymmetry because positions of weak transitional probabilities need not correspond to the clausal structure of sentences. If we take the function words which are mainly produced by the sequencing system out of the equation, the strongest competition occurs at the relative clause-initial determiner and the oblique verb. These positions correspond to the boundaries of the embedded clause. Following the argumentation of Holmes (1988) this suggests that the model might plan in units which are roughly the size of a simple clause. At the beginning of the relative clause, for example, the model is uncertain about the grammatical type of the embedding in that it activates the action role 1A as well as the thematic role 1Y of the subject NP. Statistically speaking, both continuations are equally likely given the model's linguistic experience, and the sentence initial subsequence DET NOUN THAT provides no probabilistic clue regarding the intended type of embedding. Nonetheless, the model produced the correct object-relativized embedding for all 100 tested sentences in all ten model subjects. This behavior suggests that it could not have relied on transitional probabilities but might have planned ahead in some way, based on its message input. An alternative explanation would be that the model is using its message input strictly incrementally, feature-by-feature, whenever the relevant piece of information is needed in a sentence position. For two reasons, however, this is not quite plausible. First, there are several points in production at which the model appears to 'look ahead' from the current position by activating the subsequent thematic role, for instance the 1Y role at the pronoun, the 0A action role at the preposition, and the patient role 0Y at the transitive verb. The anticipation of roles outside the current phrase indicates that the model's planning units might be larger than individual roles. Secondly, syntactic alternations in the language invert the order of roles to be sequenced at the WHERE-layer. In order to begin the active transitive clause of Figure 5.6 with the correct role, the model must attend to the activation levels of the corresponding participant features in the message ahead of production. Since the alternation bias is encoded relationally, both arguments of transitive verbs are involved in the sequencing of thematic roles in such clauses. Responding appropriately to this bias can be interpreted as clause-level planning.

The Dual-path model does not implement an explicit planning mechanism and assigns thematic roles incrementally. Nonetheless, I believe that the divide between incremental processing and hierarchical planning can be reconciled in the model. Due to its simple-recurrent architecture and word-to-word prediction mode, the model is sensitive to transitional probabilities between lexical items and it selects thematic roles incrementally based on information about the semantic value of previously produced constituents in the CWHERE-system. Due to its complete message input in the event semantics the model can engage in hierarchical planning from the beginning of production. Jointly, these properties create junctures of role competition in the WHERE-layer whose location at the beginning and end of an embedded clause is consistent with the behavioral predictions of processing accounts which favor clauses over words as the basic units of planning.[15]

[15] There is a caveat here, though. Competition points in Figure 5.6 are representative of this particular structure only. There might be different points of uncertainty in different constructions. Local uncertainty

5.2.3 Phrasal categories and argument structure

In Section 5.2.1 it was argued that the Dual-path model developed representations of lexical categories which were traceable at the COMPRESS-layer of the sequencing system. I will now examine more closely how the model represents aspects of the syntax of constructions from the input language. This analysis is restricted to simple-clause utterances, the case of relative clause constructions will be dealt with in Section 5.2.6. Specifically, I was interested in how the model partitions HIDDEN-layer space to internally represent the constituent structure of utterances. The types of these partitioned regions and their spatial relations might indicate in which way the model is chunking constituents into larger syntactic units, for instance, whether it acquired intermediate phrase structure such as verb phrases or prepositional phrases. It will be shown that the model developed representations of verb-argument structure of basic constructions but not traditional phrasal categories.

Linear discriminant analysis

The classification technique for probing the internal representations of the model used in the following sections is linear discriminant analysis (LDA for short). LDA is a statistical tool for object classification based on features. The basic idea is to divide a set of objects which are characterized in a high-dimensional feature-space into classes which are as distinct as possible. Classification is achieved by inserting hyper-planes into feature-space which separate a set of labelled 'training' items into groups. These hyper-planes are linear combinations of features ('discriminants') which describe the trained objects. The discriminants obtained in this way yield a separation of feature-space on the basis of which the class membership of novel objects can be predicted. Conversely, LDA can be interpreted as a clustering method. Objects which do not separate well in terms of their features form a cluster of similar objects.

5.2.4 Verb, noun and prepositional phrases

First, I wanted to determine whether the HIDDEN-layer represents knowledge about groups of word classes which corresponds to phrasal categories. Informally, I define the *coherence* of a group of word classes as the amount of misclassification of novel items from this group *within* the group. High coherence suggests that the model is clustering word classes from a group into a larger syntactic structure. For example, if the model represents an abstract notion of verb phrase we would expect to find that the LDA yields high coherence within the group consisting of verbs, auxiliaries, inflectional morphemes, and participles. If on the other hand the LDA strongly separates the corresponding word classes or misclassifies many items into classes outside this group of

might be strongly related to the nature and frequency of competing structures in the language and not necessarily reflect universal planning units.

constituents, there is weak coherence and consequently the model maintains no abstract knowledge of verb phrases.

In order to assess whether phrasal categories were represented in the model, I recorded the HIDDEN-layer activation states of a fully trained model subject while it was producing a set of test sentences. The HIDDEN-layer had 80 dimensions which corresponded to the features that characterize objects in LDA. In terms of sentence accuracy, the model subject tested 100% correct on all sentences in this set, thus no potential misclassification in the LDA could be attributed to production errors. The LDA training set consisted of 1.000 lexical items, the word classes and their distribution are shown in Table 5.7.[16] Then, 500 novel activation vectors were classified using the obtained linear discriminants. The overall LDA training accuracy was 93.3% correct classification, the testing accuracy 90.8% which is both significantly above chance: when the training set was used with randomized class tags the classification accuracy dropped to 17.1%. Within verb phrases, however, accuracy reached only 69.1% and all miscategorized constituents were classified within the group of word classes which occur in verb phrases. In addition, all of the total misclassification of test items involved VP constituents. This indicates that there was higher coherence within the group of word classes which form verb phrases than outside. It is therefore warranted to infer that the model has developed representations of verb phrases (narrowly defined, as consisting of verb stems, inflectional morphemes, auxiliaries and participle constructions). VP coherence can be visualized by plotting the test data in terms of the first two linear discriminants (see Figure 5.5). VP constituents all cluster around the lower left corner. The significance of this clustering should not be overstated, though, because high VP coherence arguably does not reflect the model's extraction of functional dependencies among verb phrase elements. Rather, coherence within verb phrases most likely stems from the fact that elements from different word classes can occupy similar sequential positions inside a verb phrase as illustrated by these examples:

AUX	BEING	BY	DET	ED	ING	NOUN	PAR	PER	SS	TO	VERB	WITH
62	8	15	262	21	46	262	15	116	34	26	116	17

Table 5.7: LDA training classes and their frequencies.

(4) a. ...the man kick -s a dog...
b. ...the man is kick -ing a dog...
c. ...the man was being kick -par by a dog...

Positional overlap creates higher transitional uncertainty within VP structures than in other phrases, e.g., noun and prepositional phrases, and might explain the weaker LDA-separation of these word classes. This becomes apparent when looking at noun phrase

[16] Increasing the size of the training set to up to 50.000 items had no significant effect on the prediction accuracy, and cross-validation showed that this training set was a good predictor.

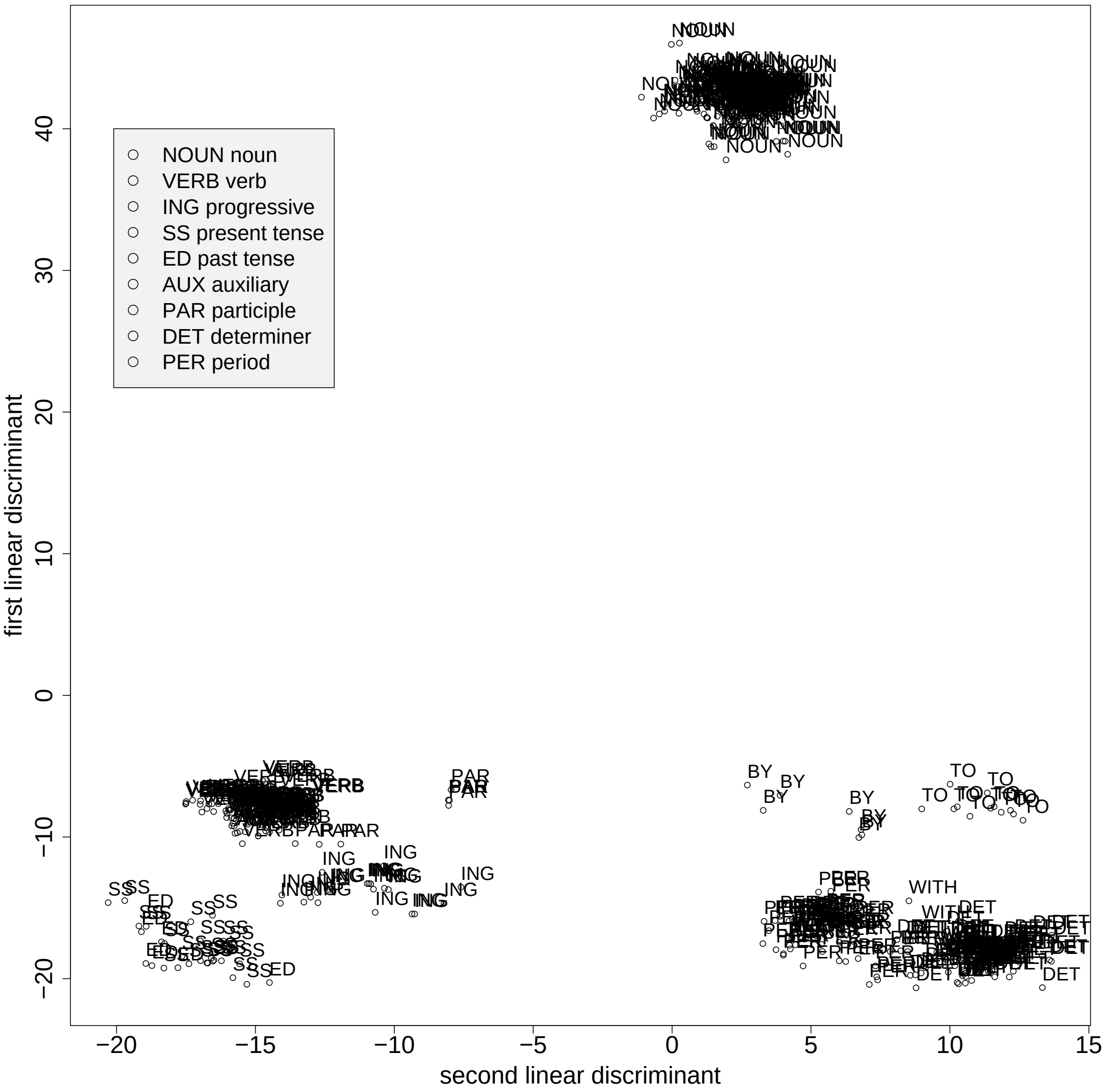


Figure 5.5: Linear discriminant analysis plot of VP coherence.

constituents in Figure 5.5. Determiners and nouns are completely LDA-separable and there is no spatial proximity in the clustering. Determiners are clustered in the bottom right corner whereas nouns are clustered in the top middle of the graph. Hence the model has not developed an abstract representation of noun phrases as a larger grammatical unit. As a consequence, there can also not be a notion of prepositional phrase represented at the HIDDEN-layer, because prepositional phrases take noun phrases as complements. Nonetheless, I took a look at the prepositional phrases which occured in the training language to see whether the model was sensitive to aspects of their similarity structure. An LDA was performed for 1.000 three-word sequences of the form

(5) a. `...to the man...`

b. ...by the boy...
c. ...with a cat...

There was 100% correct classification of these types of phrases in the LDA. Since all objects were animate nouns, the separation might be due to differences in the activation states when processing prepositions and not due to specific knowledge about individual nouns in the complements. In order to explore whether the model maintained a concept of preposition despite this separation, I conducted a cluster analysis of the data. Cluster analysis is an analytic tool (very similar to LDA) to group objects in such a way that the degree of association between two objects is maximal within groups and minimal outside. In the absence of larger phrasal units a concept of preposition should be visible qualitatively if prepositions are clustered together. They fall into distinct and spatially separated clusters otherwise. Figure 5.6 shows that the former hypothesis was confirmed. Prepositions were grouped into adjacent positions within the same superordinate cluster. The compactness of clusters and their distinctness from neighboring clusters provide measures for the robustness of this categorization. All preposition clusters were very compact because the vertical distance between the leaves and their joint root node was small.[17] This indicates that all elements of the preposition clusters were represented very similar to one another. Although all instances of by were in a distinct cluster from the instances of with, the two clusters were not very distinct because of the short distance marked with a star (*). On the other hand, the prepositions to were not very distinct from the noun cluster (marked **) and therefore associate less tightly

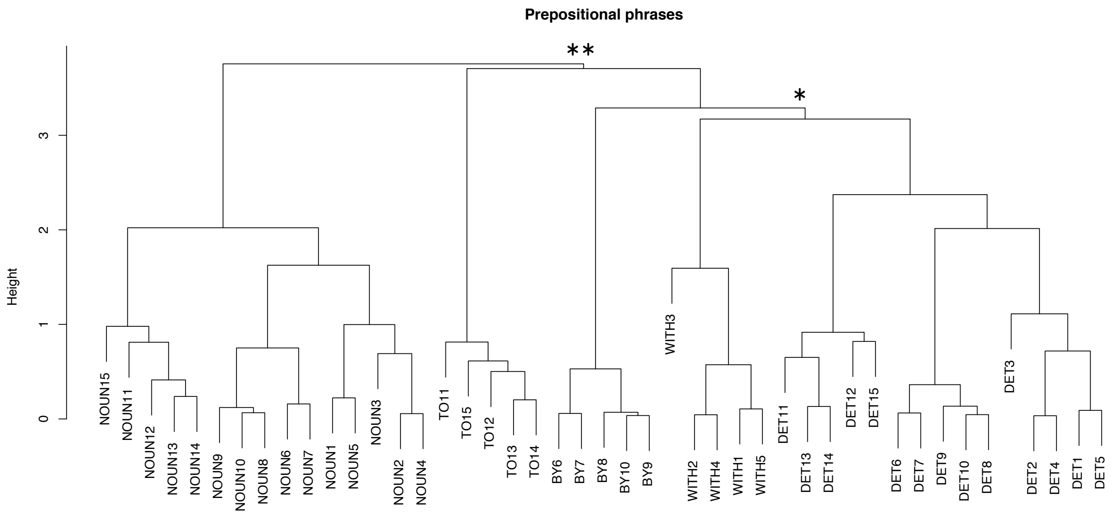


Figure 5.6: Hierarchical clustering of prepositional phrases. Nouns, determiners, and prepositions are indexed to mark different occurrences.

[17] With the exception of one outlier (WITH3).

with the other two prepositions. Despite this lack of distinctness, Figure 5.6 suggests that the model represented the functional similarity of different prepositions by spatial proximity in HIDDEN-layer space.

It is important to point out, however, that cluster analysis is merely a means of visualization, not an explanation of a discovered taxonomy. Prepositional phrases in the input language shared a number of structural similarities which may account for the proximal clustering of prepositions. All prepositions occurred post-verbal and immediately preceded the clause-final NP. Being function words they were not represented in the message-lexical system but had to be activated by the sequencing system. Their rigid sequential position may therefore have facilitated the development of similar internal representations. On the other hand, there were three participants in prepositional datives, as opposed to two participants in passives and obliques, and activating the preposition at the appropriate sentence position was complicated by the double-object alternation. Both aspects create dissimilarities with the other prepositional phrases and this may explain why the clustering is unstable with respect to the preposition `to`. Finally, prepositional phrases were quite frequent in the training corpus, so the model could draw on a lot of experience to differentiate properties of the constructions involving prepositional phrases. It can be speculated that with more structural variety in the input there would have been more pressure to evolve representations that reflected functional similarities among phrasal constituents.

The preceding analysis indicates that the Dual-path model did not developed robust internal representations of phrasal categories in the sense of traditional theory of syntax. I found some coherence among verb phrase elements, volatile clustering within prepositional phrases, and perfect separation within noun phrases. Yet, the model acquired a structurally complex target language with relative clauses and substantially generated beyond its immediate linguistic experience. What this suggests, if anything, is that the representation of phrasal categories is not required to successfully transduce between meaning representations and grammatical sentence forms. When linguists describe syntactic categories such as phrases in their favorite representational medium, such as trees or labelled bracket notation, they make no claim about the psychological reality of these particular representations. According to Jackendoff (2002), however, by putting NPs into a syntactic category it is claimed "that words group hierarchically into larger constituents that also belong to syntactic categories" and that this grouping "must be reflected somehow in neural instantiation" (p. 24). The "linguistic state-space" of the brain must encode "significant groupings of dimensions that can in functional terms be referred to as [...] syntax" (p. 25). In the Dual-path model I did not detect such groupings (e.g, for noun phrases). This suggests that the syntactic categories of theoretical linguistics may not map one-to-one onto the syntactic representations the human language processor develops in learning. It is conceivable that the functional roles linguists attribute to syntactic categories have no causal correlates in the architecture and mechanisms of human language processing.

5.2.5 Argument structure

It is a central question of contemporary linguistic theory and psycholinguistic modelling how the properties of verbs, sentence forms and sentence meanings interact. Generally, verbs are regarded as prominent bearers of semantic information because they specify the type of event described by a sentence and project the event participants. On an influential account in linguistic *syntax*, verbs specify the number and types of arguments—their argument structure—on the lexical level (Chomsky, 1965). For instance the verb `kill` would typically be considered to take two arguments, a subject and a transitive object. These grammatical categories map onto the semantic categories of agent (`Tom`) and patient (`dog`) in the sentence `Tom killed the dog`. It has been argued that the correspondence of form and meaning is encoded by linking-rules which are associated with the verb (Pinker, 1989). As Bencini and Goldberg (2000) pointed out, however, this suggestion is problematic in at least two ways. First, verbs can occur in a multitude of different argument structure frames. For example, the prototypically intransitive verb `run` can figure in numerous other configurations:

(6) a. `Tom ran.` (intransitive action)
b. `Tom ran the show.` (transitive action)
c. `Tom ran with the dog.` (oblique action)
d. `Tom ran the car into the lake.` (caused motion)
e. `Tom ran into Jerry.` (experiencer-theme)

This assortment of construction frames for the verb `run` is difficult to handle for lexicalist approaches to argument structure. A different verb sense would have to be posited to account for each sentence meaning.

Secondly, there are systematic variations in meaning associated with different argument structure frames:[18]

(7) a. `I brought a glass of water to Pat.` (prepositional)
b. `I brought Pat a glass of water.` (ditransitive)

(8) a. `I brought a glass of water to the table.` (prepositional)
b. `*I brought the table a glass of water.` (ditransitive)

Prima facie, the dative alternation (7) involves only a small change in sentence meaning. Yet, while the prepositional dative (8-a) admits of inanimate goals, the ditransitive (8-b) rarely does. Generally, there are more restrictions on the goal/recipient role of the ditransitive construction than the prepositional dative construction. To capture these systematic differences lexicalist approaches would again have to stipulate a separate verb sense for each argument structure frame.

An alternative strategy to explain argument structure has been taken by Goldberg (1995, 2006). On her account, meaning is directly assigned to abstract 'argument structure constructions' which are conceived of as linguistic primitives on a par with lexical

[18] This example is due to Partee (1965), quoted from Bencini and Goldberg (2000), p. 641.

and phrasal categories. This approach still allows verb meaning to contribute to sentence meaning. In many cases, when a prototypically suitable verb occurs in some argument structure construction, sentence meaning is derived mainly from verb meaning (e.g., the verb `give` in the ditransitive 'transfer construction') because constructional meaning does not add a novel semantic aspect. Transfer is already implied by the particular verb `give`. In other cases the argument structure construction can contribute semantic facets to sentence meaning that are not obviously induced by verb meaning alone. For example the intransitive verb `run` does not signal *caused motion* in isolation. When occurring in the argument structure construction of sentence (6-d) above, however, caused motion *is implied.* It is the abstract construction frame of *X causing Y to move to Z* which conveys this aspect of meaning. The verb `run` only designates the specific manner in which motion is caused. In this way, verb meaning and constructional meaning can interact and both contribute varying shares to overall sentence meaning. This explanatory strategy towards argument structure has been called the *constructionist approach.*

By themselves lexicalist and constructionist approaches are theoretically neutral with respect to the question whether systematic correlations between meaning and form are learnable or innate. Typically, though, constructionists endorse the view that such correlations can be acquired on the basis of child-directed speech. But constructionists differ regarding the course of development and the nature of acquisition strategies. While Goldberg and Sethuraman (2004) and Goldberg (2006) focussed on the early use of abstract categorization principles to acquire argument structure generalizations, Brooks and Tomasello (1999) and Tomasello (2003) emphasized that children's constructions are initially based on particular verbs and only expand into fully abstract argument structure patterns gradually and at a fairly late stage in syntactic development.

Although construction grammar approaches to argument structure have received considerable interest in descriptive linguistics, less attention has been paid to the psycholinguistic processes that subserve the acquisition of argument structure and the cognitive representations thereof. In this section I will utilize the Dual-path model to shed some light on these issues. By analyzing the model's internal representations I will try to address the following questions:

(i) Does the model represent argument structure at all?

(ii) If so, how much of it is due to architectural constraints and message input and how much is learned?

(iii) Are both argument structure constructions and verb-specific statistical properties contributing to argument structure representations?

(iv) If (iii), what is the proportion each factor contributes?

As in the previous section, the activation states of the HIDDEN-layer in a fully trained model subject (epoch 100.000) were recorded during the production of a set of novel single-clause utterances. These vectors were tagged with the corresponding word category labels. Nouns were distinguished by grammatical role and construction type into intransitive subjects (ITS), active transitive subjects (ATS), oblique objects (OBO), and so forth. A set of 1.000 of these sentence constituents was classified by a linear discriminant analysis. Then 200 novel constituents not used in LDA-training were classified in terms of the obtained discriminants. Figure 5.7 shows the LDA-clustering for these constituents in terms of the first two discriminants. It is apparent that all sentence sub-

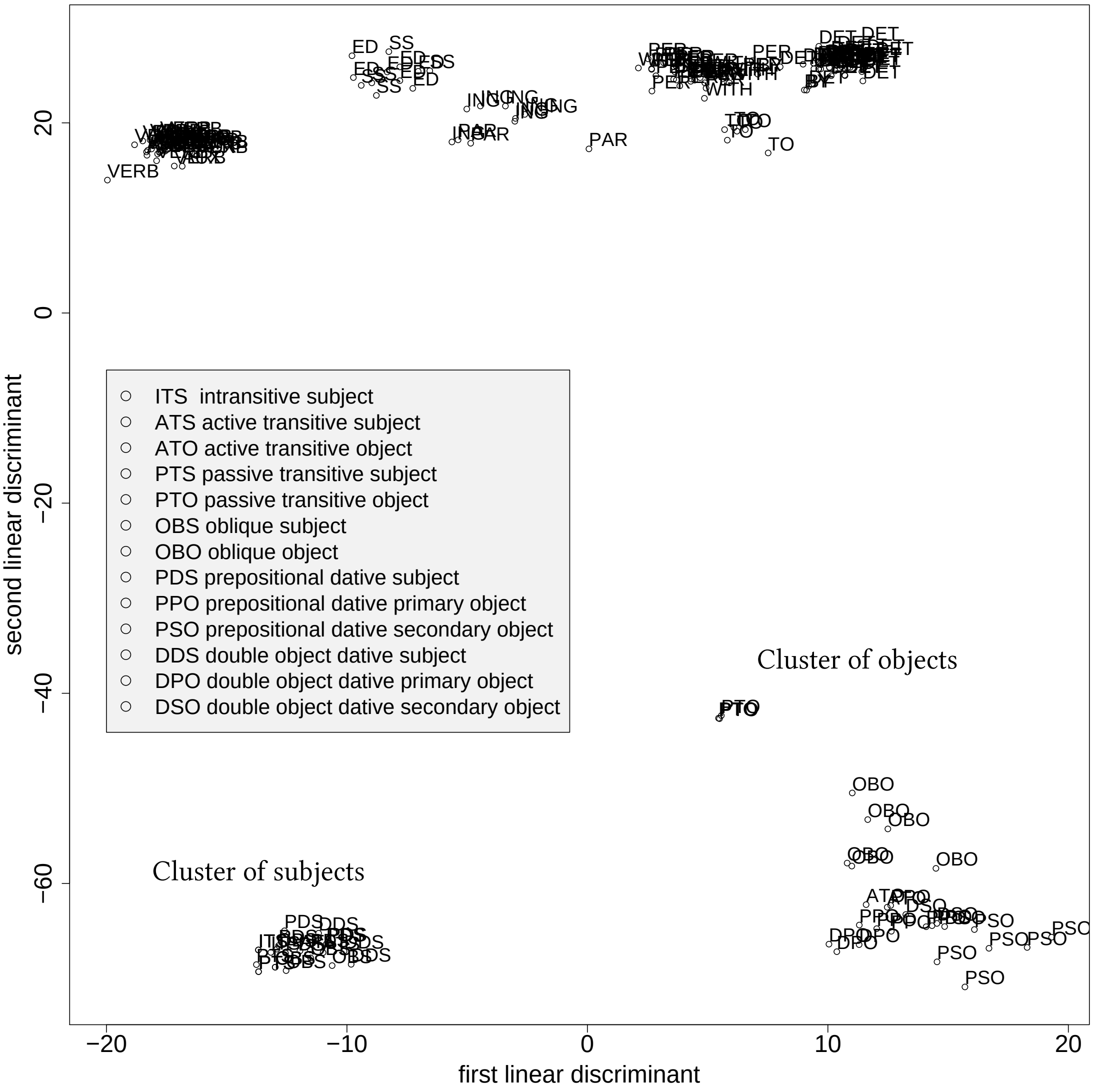


Figure 5.7: LDA of nouns distinguished by grammatical category and construction type.

jects cluster in the lower left corner of the plot whereas the different types of objects spread over the lower right corner. The overall accuracy of the LDA in distinguishing grammatical noun class was 90.8%. Specifically, the LDA reliably distinguished subjects from objects but also subjects by construction and objects by construction (see Table 5.8 on page 148 for details). The subject- and object-clusters of Figure 5.7 are shown in close-up in Figures 5.8 and 5.9, respectively.

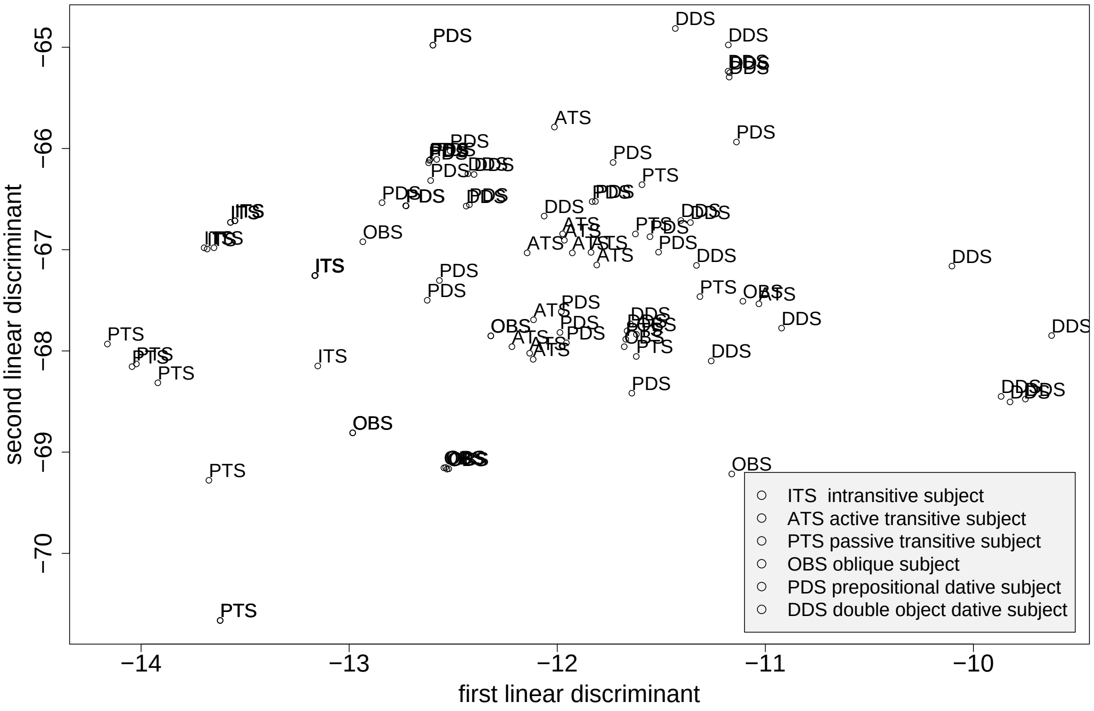


Figure 5.8: LDA zoom into subject-noun cluster of Figure 5.7 (page 145).

With the exception of inanimate dative themes, all nouns in the artificial language could occur in all argument slots across constructions. Therefore knowledge which the model maintained about the LDA-separable noun categories could not be noun-specific. In addition, because the model tested near 100% sentence accuracy on the language fragment used to obtain the HIDDEN-layer activation states, the model correctly associated verb class with the appropriate argument types in each test sentence. Hence, the clustering in Figure 5.7 was not tainted by systematic errors in verb-argument selection. For these reasons it seems warranted to conclude that the model represented the argument structure of constructions at the HIDDEN-layer.

The mere fact, however, that grammatical noun classes were LDA-separable in HIDDEN-layer space at an adult stage gave no indication whether these representations were learned and to what extent they were constrained by the message input. To clarify these issues, the same LDA was performed for three further model conditions (see Table 5.8, page 148). In the *randomized labels* condition the model received no training or message input, and the constituent labels were randomly assigned to the HIDDEN-layer vectors recorded during production of the test set. The classification score was 0% for all noun classes. Against this chance baseline, the LDA accuracy reached a total of 48.3% when the labels were assigned correctly in the otherwise identical *untrained + no message* condition. Thus, prior to training and without semantic input, the Du-

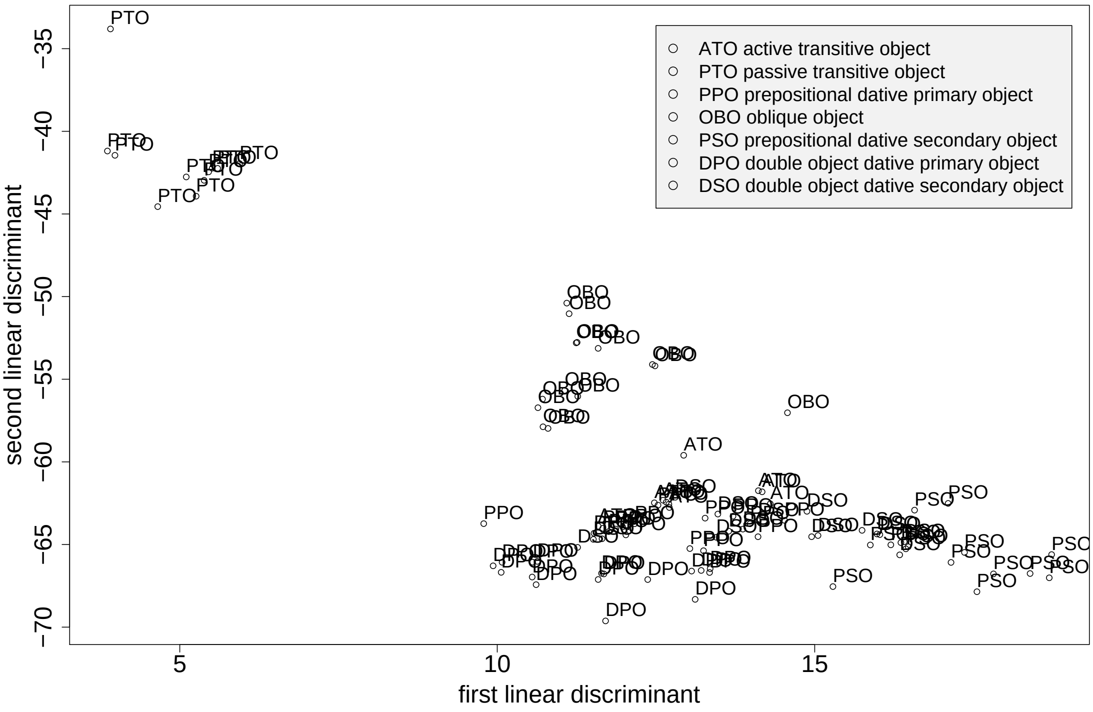


Figure 5.9: LDA zoom into object-noun cluster of Figure 5.7 (page 145).

al-path model displayed an architectural propensity towards categorizing nouns into grammatical classes.[19] When the event semantics component of the sentence message was present while recording the activation states in production prior to training, the LDA classification reached an overall accuracy of 84.3%. In particular, this message component helped the HIDDEN-layer to identify sentence subjects (compare columns 2–3, 6, 8–9 and 13 in the *untrained + no message* with the *untrained + message* condition). With the usual model training, the LDA score went up to the 90.8% accuracy reported earlier in this section. This relatively small increase in LDA accuracy when the model received training was largely due to improved classification of object types compared with the *untrained + message* condition. Consequently, the argument structure representations of the model predominantly resulted from the message input and to a lesser extent from learning. The LDA plots for the different conditions of Table 5.8 (which are not shown here) revealed that the model completely re-organized the HIDDEN-layer representational topology during learning, but grammatical noun classes were already reliably separable on the basis of constructional meaning in the event semantics. The contribution of model training to argument structure representations was most distinctly noticeable in the LDA-separation of object types.[20]

[19] I owe this observation to discussing LDA for neural networks with Morten Christiansen.

[20] Theories of argument structure are mainly concerned with object types because every English sentence requires a subject.

Noun grammatical classes[a]							
Condition	ATO	ATS	DDS	DPO	DSO	ITS	OBO
Randomized Labels	0%	0%	0%	0%	0%	0%	0%
Untrained + No Message	15.8%	0%	52.6%	42.1%	94.7%	61.9%	95%
Untrained + Message	57.8%	100%	57.8%	42.1%	94.7%	95.2%	100%
Trained + Message	84.2%	73.7%	63.2%	100%	100%	100%	100%
Condition	OBS	PDS	PPO	PSO	PTO	PTS	Total
Randomized Labels	0%	0%	0%	0%	0%	0%	0%
Untrained + No Message	0%	0%	65.4%	100%	85.7%	0%	48.3%
Untrained + Message	100%	100%	76.9%	80.8%	92.9%	100%	84.3%
Trained + Message	100%	88.5%	76.9%	100%	100%	100%	90.8%

Table 5.8: Detailed LDA statistics for the classification of novel constituents into noun categories by grammatical and construction type.

[a] ATO = active transitive object, ATS = active transitive subject, DDS = double-object dative subject, DPO = double-object dative primary object, DSO = double-object dative secondary object, ITS = intransitive subject, OBO = oblique object, OBS = oblique subject, PDS = prepositional dative subject, PPO = prepositional dative primary object, PSO = prepositional dative secondary object, PTS = passive transitive subject, PTO = passive transitive object.

What this analysis suggests is that a structured representation of constructional meaning—such as the event semantics message component in the Dual-path model—is already a very powerful predictor of grammatical relations within clauses. Human language learners face the task of constructing such a sentence message from visual and contextual information and world knowledge. This task itself is not addressed by the Dual-path production model. Yet, what the model's behavior indicates is that linking-rules between meaning and form may not have to be posited as separate *explananda* (or innate principles, for that matter) in the acquisition of argument structure generalizations. In the Dual-path model, the sentence message enforces the development of grammatical noun classes without the need of mediating linking principles.[21] Hence, the model suggests that syntactic-semantic mappings between argument structure and event structure could prove to be a by-product of meaning-to-form transduction in child language development.

It may be objected to the preceding analysis that the model's HIDDEN-layer does not represent genuine grammatical categories at all but merely reflections of the semantic content of its message input. In light of the strong event semantics influence on LDA performance this is a natural objection. There are three reasons why I consider it appropriate to refer to the clustering in Figure 5.7 (also Figures 5.8 and 5.9) as grammatical

[21] Alternatively, it could be argued that linking-rules are innate, language-*unspecific* constraints in the model since generic semantic input and architectural propensities of the model create argument structure representations through domain-general learning.

categories. First, the mapping between participant features in the event semantics (XX, YY and ZZ) and object types in the language was one-many, i.e., the same feature could map to different object types. For example, the semantic feature ZZ could map to oblique objects and indirect objects alike. Similarly, the feature YY could map to dative primary objects and transitive objects. Hence, the intended object type within a sentence was not signalled to the HIDDEN-layer by individual role features in the event semantics. Nonetheless, the LDA distinguished both pairs of object (OBO versus PSO and DPO versus ATO) perfectly in the trained model condition (table 5.8). Secondly, the mapping between semantic features in the message and subjects and objects, respectively, was also one-many. For instance, the feature YY could map to intransitive subjects (ITS) and dative objects (DPO and PSO) which were reliably separable by LDA at the HIDDEN-layer. And third, the same feature value (agent, patient, recipient, etc.) could map to different noun grammatical classes. In the active/passive alternation the agent of the action could be the subject or direct object. Since the event features in the message cut across grammatical categories in multiple ways, the representational taxonomy at the HIDDEN-layer did not reflect a *one-one* correspondence between semantic values and noun classes. Instead these representations were induced by the specific pattern of *concurrent* activation of features which jointly encoded the meaning of a construction. It was the message structure as a whole, rather than isolated features in the message, which engendered the class-separable representations displayed in Figure 5.7 (and Table 5.8). Despite being strongly influenced by the event semantics, it is therefore more adequate to view them as representations of grammatical categories rather than semantic projections.

To validate this point, I conducted an experiment which was designed to show that the HIDDEN-layer representations were sensitive to the statistical argument requirements of individual verbs. The idea behind this was to investigate whether the representations of different objects could be perturbed by purely structural properties of the main verb in the intended sentence for a given message. For two utterances expressing the same construction type, the event semantics component of the message was kept constant while the verb class was varied. If during sentence production the HIDDEN-layer representations at the object position immediately succeeding the verb remained constant across the two sentences, these representations were entirely determined by the message input and hence semantic projections (*semantic-image hypothesis*). If, on the other hand, these representations differed significantly across conditions they should rather be considered syntactic in nature because this change must be due solely to the structural properties of the verb (*argument-structure hypothesis*).

A set of test sentences was generated in which transitive verbs were placed into prepositional dative frames, and dative verbs were placed into transitive frames, e.g.,

(9) a. *`the man is teach -ing a ball to a boy .`
b. *`a dog was give -ing a mother .`

In training, the model experienced the verb `teach` only in transitive frames and the verb `give` only in dative frames. Both verb classes were disjoint in the input language. It was therefore hypothesized that the model developed distinct expectations about the types of objects that immediately follow these verbs in a sentence. A standardly trained model was tested on a set of 100 sentences of type (9-a) and (9-b), respectively, and the activation states of the HIDDEN-layer were recorded during production at the first post-verbal object position (`ball` and `mother` in the above examples). These vectors were then compared with activation states (at the same sentence position) that resulted from testing the model on a set of 100 'non-pathological', single-clause, active transitive and prepositional dative utterances encountered by the model in training. I labelled the activation states as ATO for active transitive objects, as PDO for prepositional dative primary objects, and as T\D for post-verbal objects that could either be classified by the model as ATOs or PDOs in sentences such as (9-a) and (9-b). A linear discriminant analysis was conducted to see whether T\D objects could be separated from ATOs and PDOs or whether they would cluster inseparably with either one object type. If the LDA yields the latter outcome this means that the event semantics message component has overwritten verb class-specific structural properties and the *semantic-image hypothesis* is confirmed for the HIDDEN-layer representations. If the LDA yields the former outcome this would suggest that the *argument-structure hypothesis* is confirmed.

Before reporting the results of this simulation, I will first argue that the experiment is meaningful and sound within the framework of the Dual-path model. The model architecture is symbolic in its bindings between roles and concepts in the message-lexical system. When trained sufficiently, it will sequence the action role A at the WHERE-layer at the correct sentence position and because it has learned the mapping from lexical meaning to word forms, the model can be expected to produce sentences which have novel verbs placed in familiar constructions with 100% accuracy. Consequently, possible differences in HIDDEN-layer states at the object position cannot be attributed to production errors at the verb position. Secondly, recall that active transitive patients and prepositional dative themes were encoded by the same thematic role Y. After sequencing the action role A, which produces a verb form, the model has to sequence the Y role next, no matter whether it is uttering a prepositional dative construction with a transitive or a dative verb.[22] In other words, WHERE-layer sequencing of the relevant initial segment is identical for all test sentences. Therefore, possible differences in HIDDEN-layer states at the object position can also not be attributed to different semantic affordances of the distinct verb classes. Third, the model processes sentences incrementally and each word output is fed back to the CWORD-layer which projects into the HIDDEN-layer. Thus, lexical differences between two test sentences can potentially influence the HIDDEN-layer despite identical constructional meaning in the event semantics. Fourth, because of the architecture of the model, verb semantics does not influence the HIDDEN-layer at the post-verbal sentence position. Even though the produced verb is fed back to the model and activates verb meaning in the CWHAT-layer, the CWHERE-layer filters out the verb-

[22] Or an active transitive construction with a dative or transitive verb, for that matter.

specific meaning and merely informs the HIDDEN-layer about the previously sequenced action role A. Likewise, the CWHERE2-layer will recollect only the pre-verbal thematic role which is the agent role X in all test utterances. Hence, all information that the HIDDEN-layer possibly receives about verb class is non-semantic and gathered via the sequencing pathway which is separate from the message-lexical pathway. And finally, in Section 5.2.1 above it was argued that the model learned to distinguish transitive and dative verb classes at the COMPRESS-layer. We can reasonably expect similar representations having developed at the CCOMPRESS-layer which projects into the HIDDEN-layer. In each test sentence an inflectional morpheme and a determiner occurred between the verb stem and the post-verbal object noun. However, the verb class information received from the CCOMPRESS-layer will be retained in the model's working memory—the CONTEXT-layer—over these interspersed constituents. In this way the model can still utilize verb class information at a later stage in sentence production. To sum up, the sketched experiment is meaningful because possible effects of verb argument structure on HIDDEN-layer object representations are observable in principle, and it is sound because such effects cannot be attributed to verb semantics.

I examined ten model subjects at an adult state (epoch 100.000). The linear discriminant analysis for 100 ATO, PDO and T\D items each resulted in 100% separation for all constituents in both conditions, the 'dative verbs in transitive frames' condition and the 'transitive verbs in dative frames' condition. The HIDDEN-layer representations of an exemplary model subject are plotted in Figure 5.10 with respect to the first two discriminants. It can be observed that the first discriminant does not distinguish PDO and T\D objects but the second discriminant clearly does. There is, however, considerable overlap between T\D and ATO objects on this dimension. I investigated how the T\D objects would be classified when all word categories were used as predictors in LDA training. Averaged across ten models, 91.1% of T\D objects were classified as ATO when a dative verb occurred in a transitive frame (SD 17.3), and 86.1% of T\D objects were classified as PDO when a transitive verb occurred in a prepositional dative frame (SD 12.6). Thus, in both conditions the majority of T\D objects were LDA-categorized according to construction type, not verb class. This shows that constructional meaning imposed stronger constraints on argument-structure representations than verb category. Yet, to some extent this may be an artefact of the artificial language used to train the model in which inflectional morphemes were treated as separate constituents, which required additional post-verbal processing steps before the object noun was produced. Hence, inflections weakened the influence of verb category by diluting the model's working memory. If the learning environment had inflected verb forms in the lexicon, verb category information would presumably have a stronger effect on T\D classification. In any case, it was more relevant to the current issue that the LDA perfectly separated ATO, PDO and T\D objects (Figure 5.10). This result vindicates the *argument-structure hypothesis* which claimed that the model's HIDDEN-layer representations of object types are sensitive to purely structural properties of verbs-argument frames. On the other hand, it casts doubt on the *semantic-image hypothesis* which claimed that these representations are merely reflections of semantic information in the event semantics. Placing verbs in non-

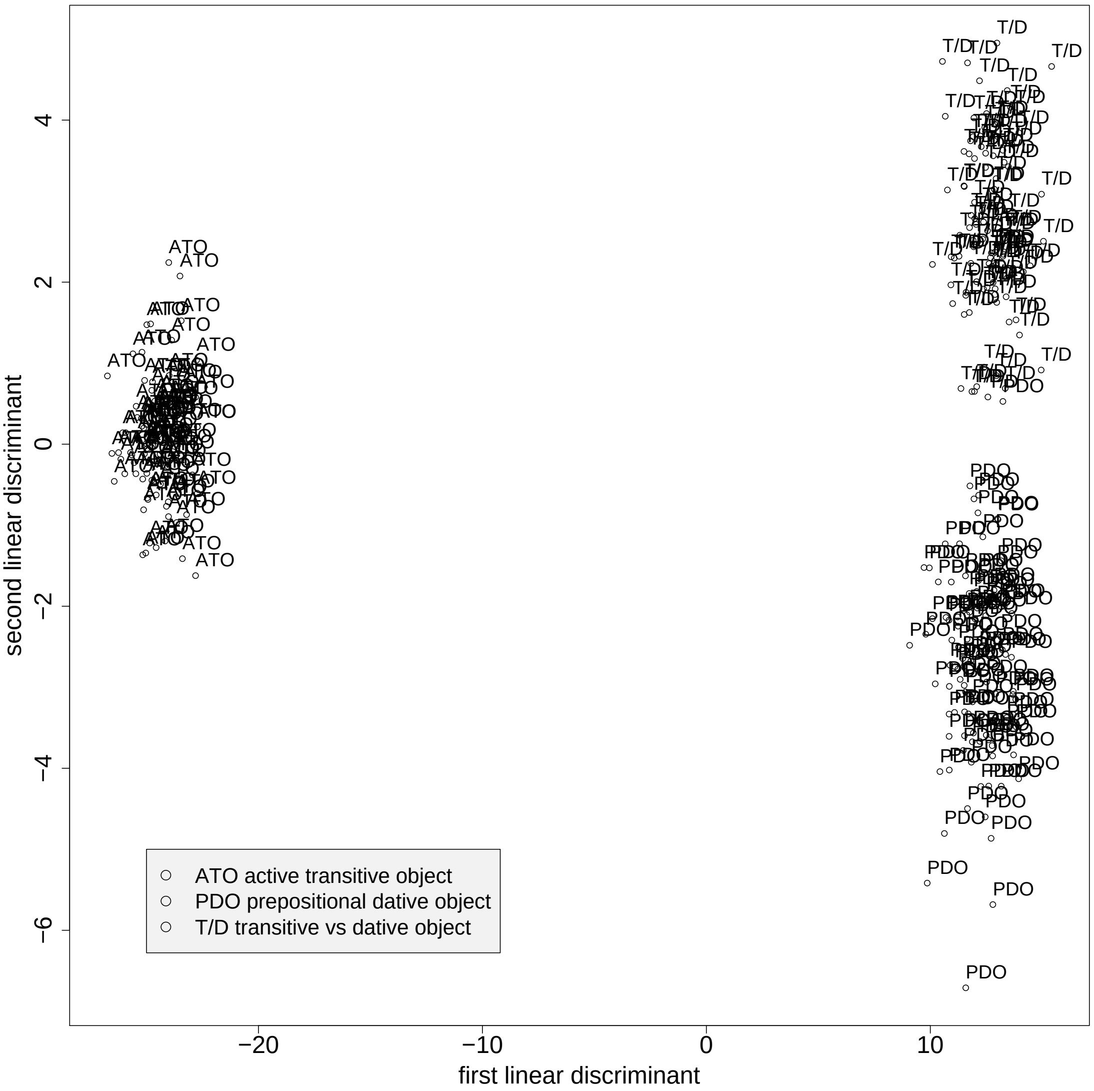


Figure 5.10: LDA plot of object nouns for transitive verbs in prepositional dative frames (T\D).

matching construction frames altered the expectations of the model for the post-verbal object type and this made T\D objects completely separable from the object types that were predicted by constructional meaning. This is a purely syntactic effect which cannot be attributed to the model's event semantics. The altered HIDDEN-layer representations solely derive from the statistical expectations of the model based on verb-class/object-type co-occurrence in the training language. Therefore, it is justified to assert that the Dual-path model maintains knowledge about syntactic argument-structure relations.

One of the most important tasks language learners have to accomplish is to distinguish verbs and nouns and identify verb-argument structure patterns. In the Dual-path model this process is driven by the meaning of argument-structure constructions in the sense of Goldberg (2006) and by the statistical regularities in verb-argument patterns. But because the model receives the complete constructional meaning from the onset of learning (situated comprehension), it is difficult to make claims about argument structure development. In fact, even at very early epochs (5.000–20.000), LDA-classification yielded results very similar to those reported for an adult state in this section. In order to determine whether argument structure generalizations are acquired on a verb-by-verb basis or by means of abstract categorization principles, we would have to model the development of semantic representations as well. It can be speculated that as semantic content becomes more and more structured towards full constructional meaning the influence of the message on the model's argument structure representations will gradually increase over the influence of verb-specific statistical information in the course of learning.

5.2.6 Clause-level analysis

In the preceding section, I looked at argument structure representations in single-clause utterances. It was tacitly assumed that representations of complex sentences would be combinatorial in the structure of simple sentences. It is quite possible, however, that the model represents the syntax of complex sentences in vastly different ways. In this section I analyze the representational similarities between different types of clauses and complex sentences. The aim was to determine the relationship between simple-clause and relative clause processing and to find out how the model represented structural properties of complex sentences, such as the subordination of clauses, attachment and relativization.

Principal components analysis

Ideally, we would like to directly inspect the internal representations of a trained neural network in order to figure out how the model solves a particular computational task. However, we cannot easily visualize these representations because each unit in the HIDDEN-layer adds a dimension of information. Apart from this complication there is no guarantee that HIDDEN-layer dimensions pick out representational dimensions that are relevant to the model's solution of a task. In fact, it is the nature of distributed representations that they cut across such dimensions.

One way of dealing with this dilemma is principal components analysis (PCA). In essence, PCA is a coordinate transformation of experimental data (i.e., HIDDEN-layer activation patterns) that (a) identifies dimensions of largest variance (and hence explanatory value) and (b) can be used to compress and visualize data by reducing its dimensionality. A data set for PCA consists of m measurement points represented by n-dimensional vectors. Each vector is normalized (by subtracting the mean), and the $n \times n$ covariance matrix of the normalized data set is computed. The eigenvectors of this matrix are the principal components, the size of the corresponding eigenvalue determines the rank of the principal components. The ratio of an eigenvalue to the sum of all eigenvalues yields the percentage of variance accounted for by the corresponding principal component. Multiplying the transposed normalized data point with the transposed principal component yields the transformed coordinates of this data point. PCA is a technique similar to LDA, however PCA is not used for classification and is not being trained on a set of data points whose class membership is known to the experimenter.

To apply PCA, a Dual-path model subject was trained as usual for 100.000 epochs on 10.000 sentences, 80% of which were single-clause utterances, 20% of which contained a relative clause. The weights of the trained model were frozen and the model tested on its entire training set. During testing, the activation vectors of the HIDDEN-layer were recorded word-by-word for each sentence in the training set, yielding a data set of roughly 99.700 vectors in 80 variables. A PCA was conducted on this data. Figure 5.11 shows the ordered size of the eigenvalues of all principal components, Table 5.9 summarizes the cumulative variance explained by the first 16 principal components. Together these components accounted for 80% of all variance in the described data set.

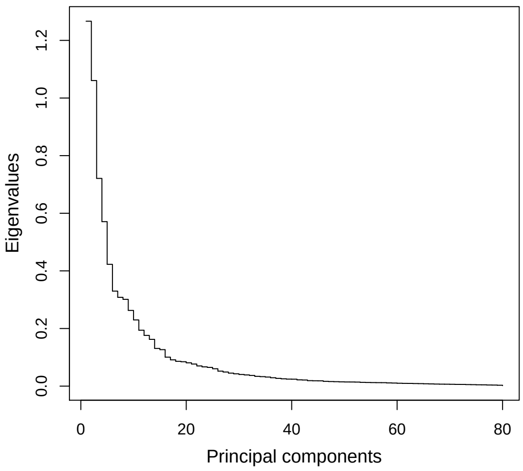


Figure 5.11: Size of the eigenvalue of each principal component.

Principal component	C1	C2	C3	C4	C5	C6	C7	C8
Standard deviation	1.13	1.03	0.85	0.76	0.65	0.57	0.56	0.55
Proportion of Variance	0.16	0.13	0.09	0.07	0.05	0.04	0.04	0.04
Cumulative Proportion	0.16	0.29	0.38	0.45	0.50	0.55	0.58	0.62
Principal component	C9	C10	C11	C12	C13	C14	C15	C16
Standard deviation	0.51	0.48	0.44	0.42	0.40	0.36	0.36	0.32
Proportion of Variance	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01
Cumulative Proportion	0.66	0.68	0.71	0.73	0.75	0.77	0.78	0.80

Table 5.9: Cumulative proportions of variance explained by the first 16 principal components.

Clause type comparison

We would like to know how the model encoded structural properties of complex sentences such as distinct clause types, clause level, attachment and relativization. To probe these aspects of sentential structure, the model was tested on two novel sentences

(10) a. `the nurse push -s a cat .`
b. `the nurse push -s a cat that was sleep -ing .`

and HIDDEN-layer activation was recorded for both sequences. During processing, both sentences describe a trajectory through hidden unit space which represents an admissible grammatical construction vis-à-vis the model's experience. In Figure 5.12 these trajectories are plotted in the coordinates of the first two principal components. The clausal overlap between both sentences follows very similar trajectories in terms of these coordinates. Both noun phrases of the two transitive clauses (`the nurse` and `a cat`) are represented in almost identical regions of HIDDEN-layer space. The main clause verb form `push -s` on the other hand is represented differently in both sentences. In the complex sentence the verb positions are slightly shifted along the first principal component compared with the single-clause utterance, yet both sentences describe qualitatively identical trajectories. This suggests that the model is processing simple-clause utterances in a similar way as main clauses of right-branching constructions and indicates that the similarity structure of both sentence types is reflected in the representations the model developed during learning. It does not treat constructions of distinct clausal complexity as separate entities but rather builds complex structure from simpler clausal units. The model has acquired a notion of *clause* and therefore a notion

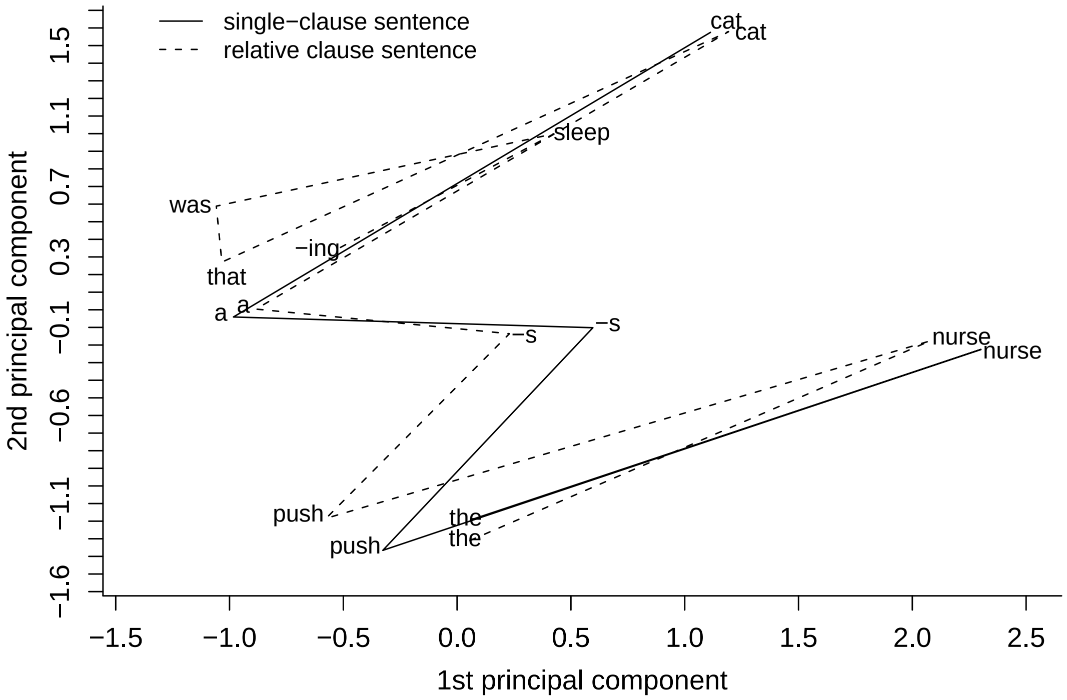


Figure 5.12: HIDDEN-layer trajectories of lexically identical single-clause sentence (10-a) and main clause of complex sentence (10-b).

of combinatorial syntax.

Some caution should be issued here, though. The purpose of LDA was to separate word category representations into distinct classes. When plotting results in terms of linear discriminants, information was lost. This was not problematic because the critical information was in the quantitative LDA-separation statistics, not in the graphs themselves. PCA on the other hand is used here to determine representational similarity among structural types from data visualization. But how do we know, for instance, that noun representations in Figure 5.12 are not separated by other principal components? Plain and simple, unless we inspect all remaining 78 principal components we cannot rule out this possibility. Since the first two principal components explain more variance in the data, however, we know with certainty that any noun separation along lower principal components will be less distinct. Thus we can conclude that the two clause types in Figure 5.12 are *mainly*, but perhaps not exclusively, distinguished by the model in terms of verb form representations. And secondly, we can make PCA more significant by looking at the same two principal components in all analyses. If similarities persist or disappear across conditions this will render individual PCA plots more meaningful and indirectly inform us about representational differences.

To illustrate this point, consider the two sentences

(11) a. `the nurse that push -s a cat was sleep -ing .`
b. `the nurse that a cat push -s was sleep -ing .`

which share the same main clause, attachment site and the same lexical material in the subordinate clause, where (11-a) is subject-relativized and (11-b) object-relativized. Figure 5.13 shows the hidden space trajectories of the two sentences, again in terms of the first two principal components. Here we observe virtually no positional difference between the two main clauses (`the nurse was sleep -ing`) on the verb form. Unlike in Figure 5.12 both main clause trajectories are almost congruent. Hence, the model represents the difference between single-clause utterances and main clauses of complex sentences in the verb position but does not make such a distinction between main clauses of constructions with different relative clause types. This suggests that the main clause representations in center-embedded structures are not influenced by the intervening relative clause material, indicating that the model is organizing complex sentences into autonomous clausal units. The representational difference between subject- and object-relativized clauses, on the other hand, is rather large. Immediately following the pronoun, both relative clause trajectories diverge into different regions of hidden unit space. The embedded NPs (`a cat`) which assume distinct grammatical roles are spatially separated, and even more so the embedded VPs (`push -s`). This larger representational difference on the VP makes sense in terms of argument structure. While `cat` is an argument of `push` inside the relative clause in both sentences, in sentence (11-a) `the nurse` from the main clause is the agent of `push` whereas in (11-b) it is `a cat` from the relative clause, and likewise for the patient of `push`.

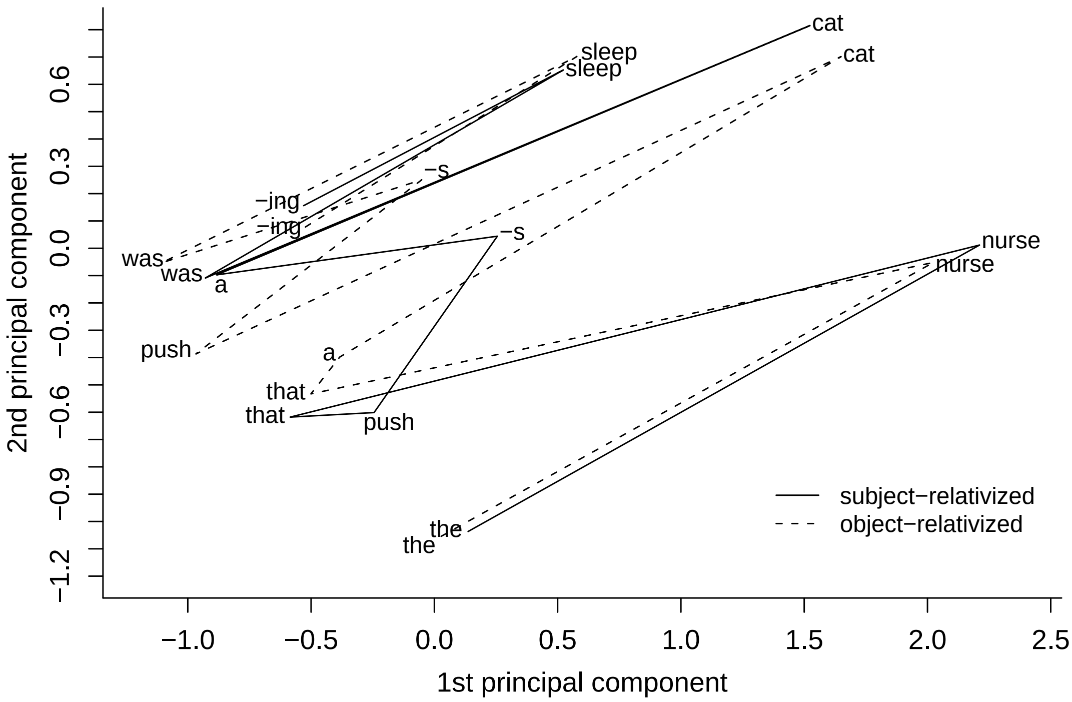


Figure 5.13: Subject- and object-relativized clauses in sentences with identical main clause, (11-a) and (11-b).

To determine the similarity structure between main and embedded clauses I looked at complex sentences in which the relative clause was lexically and sequentially identical to the main clause, and in which the arguments of both verbs were co-referential, such as

(12) `the nurse that push -s a cat push -s a cat .`

Figure 5.14 shows the trajectory of sentence (12). The relative clause follows a trajectory which is qualitatively similar to the main clause but is phase shifted in HIDDEN-layer space. The trajectory of the relative clause is iterated for the main clause but in a different region. This indicates that the model's internal representations distinguish clause level on all clause components, not just the verb phrase. Figures 5.12 and 5.14 jointly suggest that the model experiences main clauses of complex sentences with a right-branching relative clause as more similar to single-clause utterances than interrupted main clauses to lexically identical relative clauses. It also suggests that the attachment site may have a critical influence on the way clauses are represented.

I examined this issue by analysis of the following two sentences:

(13) a. `the nurse push -s a cat that was sleep -ing .`
b. `the nurse that was sleep -ing push -s a cat .`

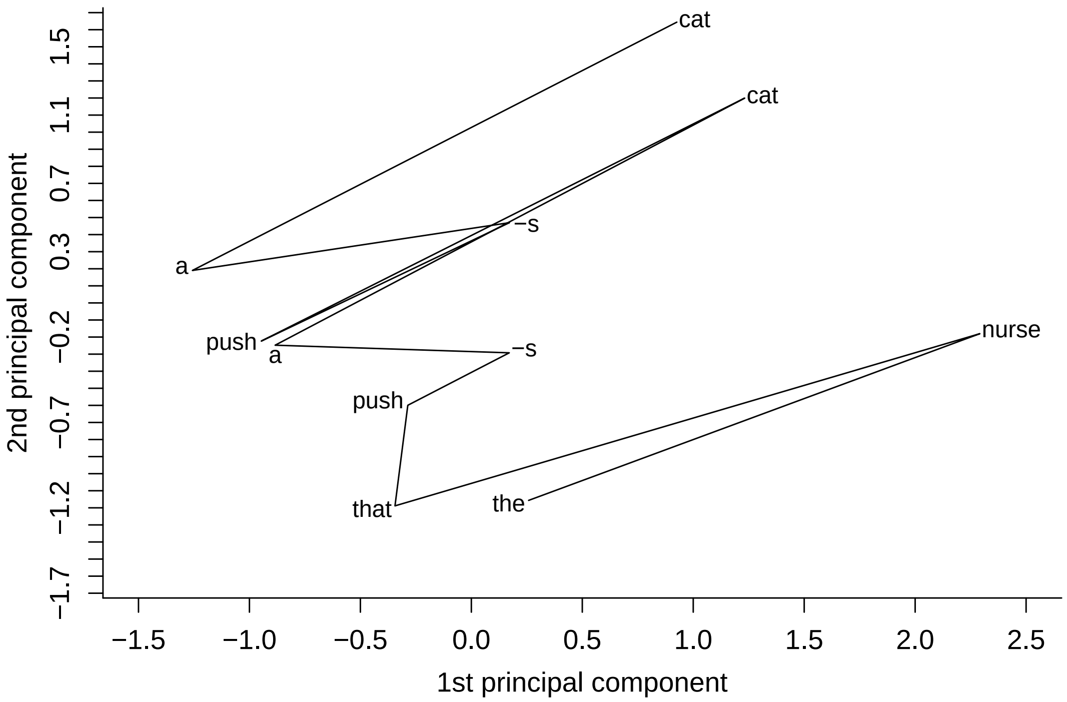


Figure 5.14: Complex sentence with identical main and subordinate clause.

Both sentences share the same main clause and lexically identical subordinate clauses but in (13-a) the relative clause is attached to the main clause object while it is attached to the main clause subject in (13-b). Thus, the sentences differ in terms of the position, grammatical role and lexical content of the head noun. The principal components plot for these sentences is shown in Figure 5.15. In contrast to Figure 5.13, in which main clause trajectories were congruent, it can be observed that subject- versus object-attachment differentiates the representations of the main clause. Both main clauses are qualitatively similar (imagine a dotted line from `nurse` to `push` in the subject-attached trajectory) with noun phrases placed in roughly the same regions, but there is a considerable spatial difference between the main clause verbs. The distinction between subject- and object-attachment appears to be marked mainly on the verb rather than the head noun. This observation underlines the central role of the verb in the model's syntactic representations. It could be argued that the verb is sequentially preceded by different lexical material in both sentences, whereas the head nouns are not, and that this accounts for the verb spatial separation. Notice, however, that in Figure 5.14 both direct objects are immediately preceded by the same lexical items and there is spatial separation nonetheless. Both main clauses in Figure 5.15 start in the same region of HIDDEN-layer space, diverge on the main verb and converge again on the direct object. This behavior is in line with the findings from Section 5.2.2. The qualitative similarity of main clause trajectories in syntactically distinct structures suggests that clauses are a basic processing unit for the model. The congruence of paths for the two sentence-initial

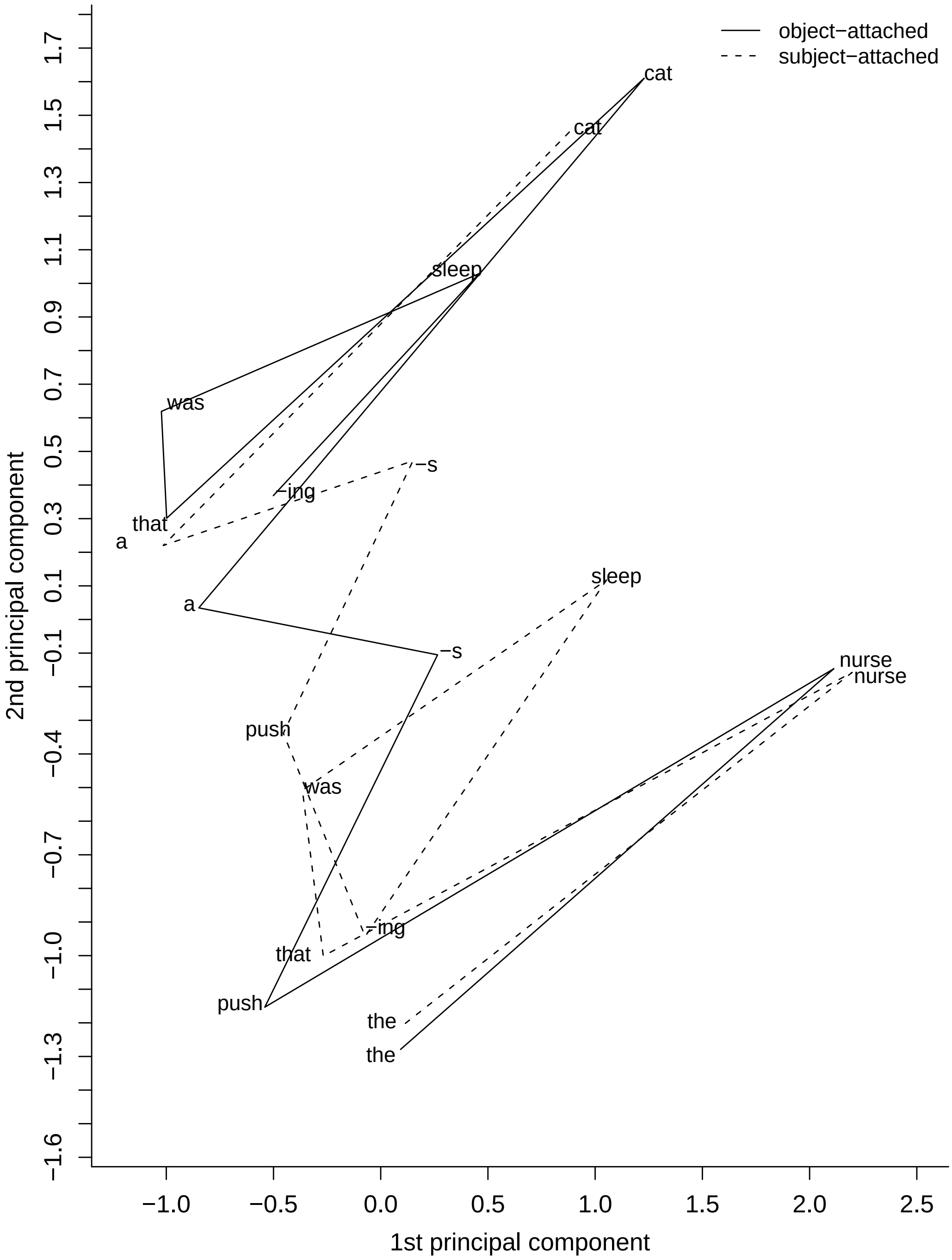


Figure 5.15: Sentences with identical main clause but object- and subject-attached relative clause (y-axis stretched for enhanced visibility).

noun phrases (`the nurse`) indicates that the hierarchical structure of the two sentences was not fully planned at the onset of processing. Differences in representations only begin to unfold incrementally where and when they become relevant in structural selection. Figure 5.15 also shows that the two lexically identical relative clauses `that was` `sleep -ing` describe very similar paths. Surprisingly, however, they are located far apart from each other in hidden unit space. This seems to indicate that the model might experience and represent center-embedded and right-branching relative clauses in different ways, as distinct clausal units. It is more plausible, however, that the dislocation is due to the fact that these clauses modify different head nouns at different sentence positions.

To summarize the results of PCA for the syntactic representations at the model's hidden-layer, I propose an ordering of similarity between pairs of clause types in Table 5.10. This ordering should be taken with a grain of salt, however, because of the methodological limitations of PCA and because it is based on the appearance of qualitative similarity of trajectories and their spatial separation. Whether these are the relevant parameters to measure *the model's perception* of similarity and dissimilarity between clause types, and if so, how they should be weighted, are very different issues. What further obviates any hard conclusions to be drawn from this analysis is the question whether similarity between representations points to facilitation or interference between similar (dissimilar) clause types in learning and processing. For instance, the representational similarity between single-clause utterances and main clauses of right-branching structures might indicate that the former structures, which are very frequent in the input, facilitate learning of the latter. On the other hand, it might indicate that since their representations are spatially close in hidden-layer space, the model has more difficulty distinguishing these two structures than others. This might create competition for structural selection and therefore sources of confusion in syntactic development. Ultimately, these issues can only be resolved through behavioral experiments and error analysis. I now turn to investigating the generalization capabilities of the Dual-path model.

	Pairs of clause types		
Dissimilarity ↓	simple	∼	uninterrupted main
	uninterrupted main	∼	interrupted main
	interrupted main	∼	relative
	subject-attached relative	∼	object-attached relative
	subject-extracted relative	∼	object-extracted relative

Table 5.10: Pairwise similarity ordering by clause type.

Chapter 6

Generalization

In Chapter 4, the model generalized to a combinatorially complete language with one embedding. In this chapter, I investigate the model's generalization capacities in more detail. I show that the model transfers lexical knowledge from one clause to another, that it generalizes verb argument structure across embeddings, that it behaves strongly systematic, and that it is recursively productive on a clausal level. I will also argue that the model's behavior on these tasks resembles human sentence processing and explain why this is the case.

6.1 Introduction

With increased structural complexity in a language, the number of grammatical constructions grows exponentially. By means of relativization, for instance, the simple-clause transitive construction alone can be combined into four distinct relative clause constructions (center-embedded and right-branching, both subject- and object-relativized). For a data-driven learning system this implies that the proportion of grammatical forms to which it is exposed decreases with more structural complexity. As a consequence, the demands on the learner's generalization mechanisms increase if the grammar of the target language is to be learned to satisfaction from sparse input. In this chapter I will argue that strong generalization can be obtained in the Dual-path model because novel structures can be built from semantic similarities shared with experienced structures.

6.2 Knowledge transfer

In the artificial language I have been studying so far, multi-clause utterances were composed of single-clause constructions basic to human experience. In order for the Dual-path model to be able to generalize syntactic knowledge to novel complex constructions it is a prerequisite that knowledge transfers from one clause to another. The experiment I describe in this section is designed to test the *transparency* of the model's syntactic representations. There are two hypotheses to be evaluated here. First, the model might develop clause-specific representations, depending on whether a basic construction occurs in a simple sentence, in the main clause, or in an embedded clause of a complex sentence. If syntactic representations are separable in this way, syntactic knowledge will not transfer between clauses and it is unlikely that the model generalizes substantially beyond linguistic experience. Alternatively, the model might develop shared representations for basic constructions, irrespective of the locus of occurrence in the training sentences. In this case, we have reason to expect that the model might be capable of structural generalization, although it would be a further open question whether it can access and utilize these representations in processing novel constructions. I tested these hypotheses for verb-argument structure information in transitive frames in a simple experiment. I used the language described in Chapter 4 which generated utterances with at most one relative clause and the model received TOPIC-FOCUS message input in training. In contrast to Chapter 4, however, I fixed the order of events in the EVENT SEMANTICS-layer so that semantic features of embedded clauses would not be trained in simple-clause processing. This was to exclude possible transfer resulting from feature overlap in the message input. The model was trained on 5.000 simple-clause sentences interleaved with 5.000 relative-clause sentences. The verb `hit` occurred only in simple-clause active and passive transitive sentences, and in no complex utterance. In addition, regardless of tense or aspect marking, the verb stem `hit` was always immediately followed by the nonce word guu, as in the sentence:

(1) `the man was hit` *guu* `-ing the dog .`

The model develops representations of abstract syntactic frames in the sequencing system and the experiment aimed at testing the transparency of these representations. The nonce word guu was not represented in the message-lexical system, so that the model could not rely on semantic information to produce it. In this way, it had to learn to sequence guu in the `hit` verb phrase (but not in other verb phrases) without drawing on semantic cues. The model was then periodically tested on 200 multi-clause utterances which contained a transitive relative clause with `hit` as the subordinate verb, e.g.,

(2) `the nurse that a teacher was being hit -par by walk -s .`

In the test utterances the word guu did not occur after the verb stem `hit`. The production accuracy for these utterances was compared with a control condition in which guu did not appear in any training sentence. The idea behind this design was that the model would learn in its sequencing system that the word guu was an integral part of transitive frames whenever the main verb was `hit`. If there was syntactic transfer from simple clauses to embeddings in complex utterances we should observe *disruption* in the processing of test utterances (which did not contain guu). If, on the other hand, the model learned distinct representations for simple and embedded transitive constructions we should observe no difference in production accuracy between both input conditions.

Results from this experiment are shown in Figure 6.1. The graph reveals two inter-

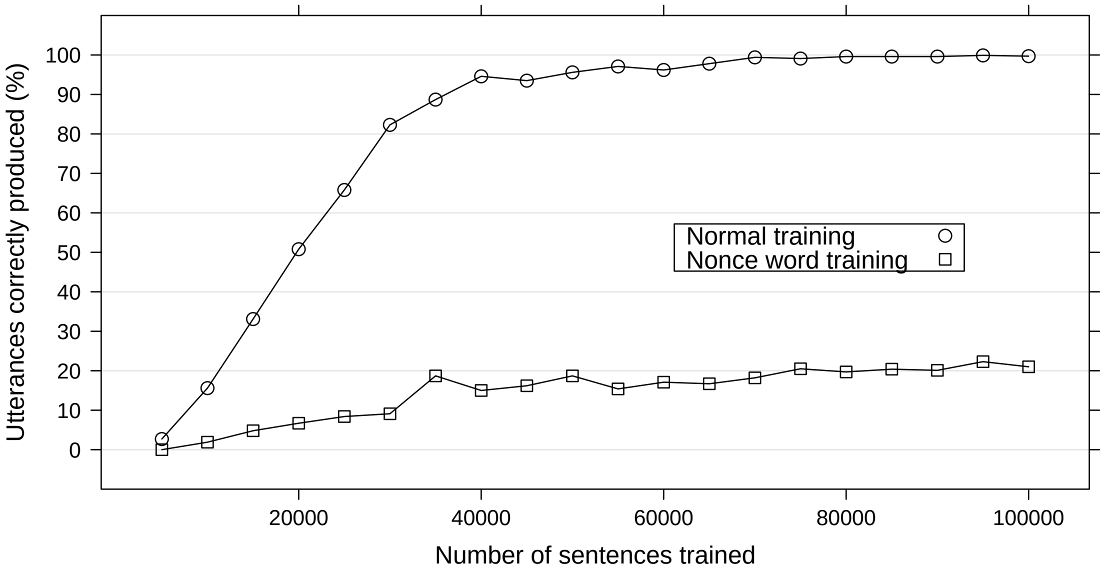


Figure 6.1: Transfer of verb argument structure between clauses as witnessed by the disruption of production in nonce word training.

esting properties of the Dual-path model. In the control condition ('normal training'), all test utterances are produced correctly at the end of training despite the fact that the model never experienced the verb `hit` in an embedded clause during learning. This behavior will be explained in Section 6.5 on lexical generalization. Secondly, in the 'nonce word training' condition, the model had considerable difficulty producing the test utterances and did not reach above 20% accuracy. The nonce word `guu` in single-clause transitive frames is severely disrupting the model's embedded clause production when `hit` is the subordinate verb. This indicates that the model developed syntactic representations for basic constructions which are not specific to the occurrence of these constructions in different clausal positions within hierarchically distinct sentences. The model's representations are *transparent* in that knowledge can transfer between clauses. We can therefore expect that learning simple clause structures supports the processing of more complex structures. Hence the Dual-path model satisfies an important precondition for all kinds of syntactic generalization that I will investigate later in this chapter.

6.3 Extension of language and semantics

For the following series of generalization experiments it was required to extend the artificial language to deeper embeddings. I will briefly describe such a language and the modifications of the message input in this section. As in Chapter 4, the artificial language consisted of basic constructions from which more complex constructions were assembled through relativization. These constructions are listed with examples in Table 6.1. By attaching relative clauses to noun phrases in any permissible way (i.e., respecting

Structure	Example
Intransitive	`the cat was sleep -ing .`
Transitive	`the woman kick -ed the teacher .`
Transitive passive	`the teacher is kick -ed by the woman .`
Prepositional dative	`a girl throw -s the stick to the cat .`
Oblique	`the nurse is play -ing with a dog .`

Table 6.1: Basic construction types in the language from which more complex sentences were built.

animacy constraints), a combinatorially complete language with one subordinate clause was formed. Three examples of such sentences are given in Table 6.2. Relative clause constructions could be center-embedded (examples (3) & (5)) or right-branching (example (4)) and could have subjects relativized (examples (3) & (4)) or objects (example (5)). Noun phrases in these structures could then be further relativized *inside* the relative clause to form sentences with two embeddings. Again the complete language with two such embeddings was admissible. Sentences with two embeddings could then be relativized once more to form all possible sentences with triple embeddings from the basic constructions of Table 6.1. These sentences with multiple, nested embeddings rapidly become very difficult to process for humans, in particular if noun-verb dependencies are not semantically constrained. Table 6.3 depicts two examples of such sentences, one with double and one with triple embeddings. More levels of embedding give rise to a

Example		Main clause	Subordinate clause
(3)	`the mother that is walk -ing was show -ing the cherry to the boy .`	Dative	Intransitive
(4)	`a boy kick -ed the man that was push -par by a dog .`	Transitive	Passive
(5)	`the cat that a father throw -ed a ball to is jump -ing with a girl .`		

Table 6.2: Sentences with one relative clause from the artificial language.

Examples of double and triple embedded sentences.	
(6)	`a dog was push -ing a mother that a woman that the brother is show -ing a milk to was being kick -par by .`
(7)	`a father that the nurse that a girl that is being hit -par by a dog present -ed the orange to kick -ed is being carry -par by a woman .`

Table 6.3: Sample sentences with multiple nested relative clauses from the artificial language.

combinatorial explosion of construction types and an exponential increase of sentence tokens in the language. The total number of construction types in this language was 6571. With a small lexicon of only 48 words, particles and inflectional morphemes, it was possible to create roughly 2.49×10^{18} different sentence tokens from this grammar.

Accommodating the model's message representation to this more complex language was straightforward. As in the case of simple-clause utterances, concepts were dynamically bound to thematic roles in the WHAT-WHERE-system. The thematic structure of each clause was represented by event features in the EVENT SEMANTICS-layer just as in the single-clause case. Different clauses were then linked in the event semantics by topic and focus features as described in the discussion of the TOPIC-FOCUS message in Section 4.3.7 of Chapter 4. Multi-clause utterances, however, required multiple topics and foci to be represented in the message. For instance, sentence (6) in Table 6.3 required two topic features, one for `mother` in the main clause, and one for `woman` in the first relative clause. Likewise, the meaning representation of (6) required two focus features in order to inform the model about the intended gap sites. As before, all event participants gapped in the target sentence were present in the message-lexical system in the form of a synaptic binding between corresponding thematic role and concept, and represented by an active feature in the event semantics. Thus, the message for utterances with three or more clauses was a direct extension of the message for utterances with one relative clause.

6.4 Structural generalization

It is beyond controversy that natural language syntax is combinatorial in the minimal sense that simple-clause constructions can be combined into multi-clause utterances. A construction such as sentence (6) of Table 6.3 is grammatical by virtue of the grammaticality of its component clauses and the grammaticality of the relativization construction. In this sentence a relative clause modifies the direct object of an active transitive clause. This relative clause is a passive transitive construction in which the subject is modified by another object-relativized prepositional dative clause. Although it is quite unlikely that the reader has ever encountered the syntactic structure of sentence (6), it is possible to process this sentence with little effort. Thus, humans have the ability to combine single-clause structures they already master into novel multi-clause constructions by means of relativization. In this way, a larger repertoire of complex constructions can be assembled from simpler units through a single procedure in production and comprehension.[1] This procedure provides a form of structural generalization because it allows the human language system to process combinations of simpler units that have not been encountered in the ambient language. It is an important question whether this capacity can be learned through linguistic experience or whether it should be assumed to be part of our biological endowment for language. Moreover, if it can be learned, it remains to

[1] Whether relativization extends indefinitely is of course an issue of much controversy, see Section 6.7 below.

be determined whether learning is accomplished by language-specific or domain-general mechanisms.

The extended language and message from the previous section enabled me to investigate these questions in the framework of the Dual-path model. In Chapter 4 the model learned a combinatorially complete language with one embedding. In these experiments, input was sparse since the model encountered only a minute fraction of all possible sentence tokens that the artificial grammar generated. The model developed abstract construction frames and the dynamic bindings in the WHAT-WHERE system helped it to correctly produce the entire target language in testing. The input was saturated, however, in the sense that the model encountered sentence tokens of most of the construction types in the language. In this section I describe an experiment in which the model was not exposed to all construction types of the target language, in particular not all constructions with three embeddings. The model was then tested on sentence tokens representing novel constructions, and on sentence tokens instantiating constructions that were encountered during learning. If the model is fully capable of structural generalization we should observe similar production accuracy on both test sets and this accuracy should be high. If the model is not capable of structural generalization we should observe low production accuracy on novel constructions. If the model is capable of some structural generalization we should observe moderate production accuracy on novel constructions. The difference between the accuracy for novel and trained constructions then provides a measure of the degree of structural generalization.

I trained the Dual-path model on 10.000 sentence tokens from the language described above. 40% of these tokens were simple sentences generated from the 5 basic construction types of Table 6.1. 30% of the sentences in training contained one relative clause, 20% contained two nested relative clauses, and 10% of the training items contained three nested relative clauses. All utterances were randomly selected. Figure 6.2 summarizes the distribution and shows the number of different construction types in the language by the number of embeddings. This distribution does not match realistic input to a human learner where sentences with two or even three embeddings hardly ever occur. Rather, the experiment was intended as a proof of concept that neural networks can develop syntactic representations for a structurally complex language and generalize this knowledge to novel constructions not attested in the input.

The first question of interest was whether the Dual-path model could at all learn the complex language to a satisfactory degree based on this distribution. Three factors could be expected to prevent this. The enormous expressivity of the input language may obstruct learning abstract syntactic frames altogether because more distinct sentence tokens entail less regularity in the language and hence weakened transitional probabilities between word categories. Secondly, more clause disruption through relativization, may impede learning to sequence thematic roles in the appropriate order in the message-lexical system. And third, more clauses entail that the model has to attend to vastly more event features in its message input. This may exhaust the model's capacity to detect activation differences between features, and to sequence clauses in the correct order.

Surprisingly, these potentially detrimental factors did not prevent the model from

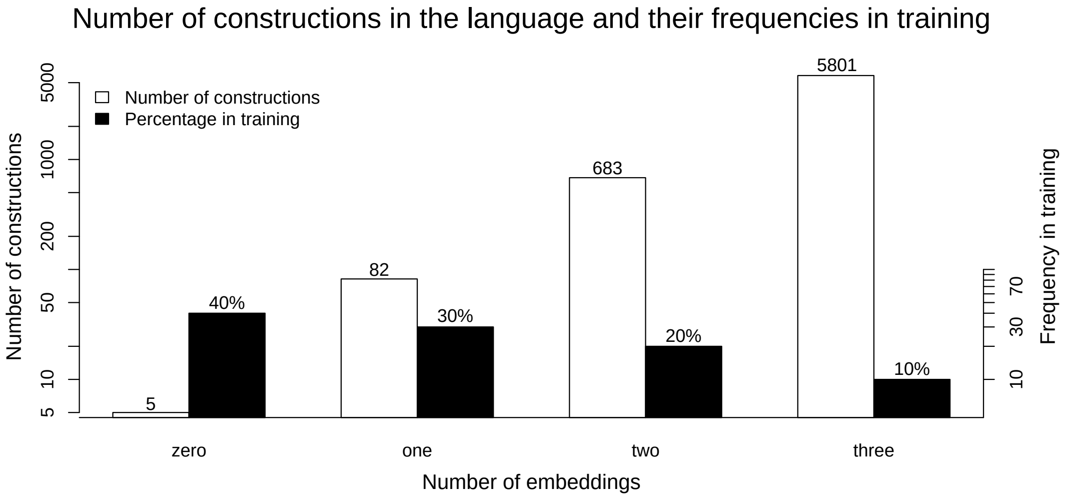


Figure 6.2: Training environment in the structural generalization task.

learning the target language to a high degree of accuracy. Not surprisingly, however, the amount of training required was considerably larger than in previous simulations. The model's learning curve for the language with three embeddings is shown in Figure 6.3 by number of embeddings. Simple clause structures were learned first, followed by sentences with one relative clause. Next are sentences with two relative clauses and sentences with three relative clauses developed slowest.[2] The language fragment consisting of sentences with zero, one or two nested embeddings was learned to perfection (>95%). Sentences with three nested embeddings were learned with roughly 65% accuracy.[3] This is no drawback because such sentences are very difficult to process for humans as well, who experience processing limits at two embeddings already (Miller and Isard, 1964).

To determine the amount of structural generalization that may have occurred at epoch 200.000, it is necessary to analyze the composition of the set of tested utterances in relation to the training set. The test set contained 200 utterances with three nested embeddings. On average, 197 of these items were instantiations of different construction types.[4] 82.6% of these unique constructions were novel constructions, i.e., the model had not seen sentence tokens representing these construction types in training. Thus, the model did not encounter the majority of tested construction types during learning. In conjunction with the overall accuracy of 65%, this indicates that a substantial amount of structural generalization has occurred. Table 6.4 depicts the model's performance for

[2] Again, all results were averaged over ten model subjects which received different randomized training set.

[3] As before, accuracy was measured in terms of a perfect word-to-word match between target utterance and actual model output.

[4] Two constructions were classified as different when they had a different clausal profile and/or a different relativization profile; differences in tense/aspect were ignored.

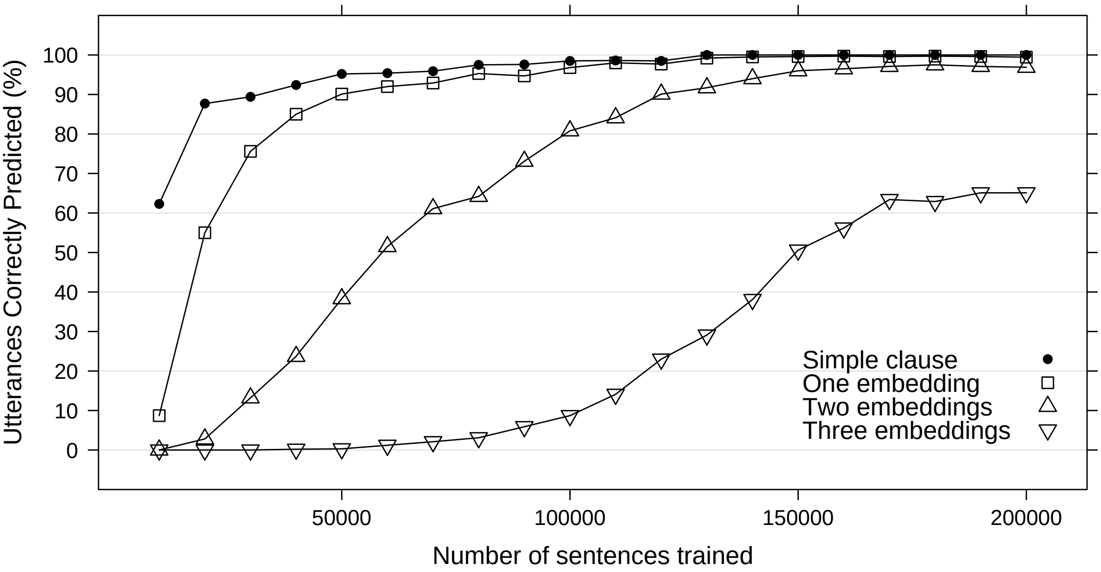


Figure 6.3: Dual-path model performance by number of embeddings.

sentence tokens, and for novel and trained construction types. A repeated-measures

Tested structure	Number	Accuracy
Total sentences with three embeddings	200	65.11%
Total unique constructions	197	65.19%
Unique constructions in training set	35	68.46%
Unique constructions *not* in training set	162	65.06%

Table 6.4: Performance on the test items in detail; figures averaged over model subjects.

ANOVA was conducted which did not detect a difference in mean accuracy (F(3,36) = 0.47, p = 0.70) on these different sets of sentences. In particular, there was no statistically significant difference in whether the model had experienced a construction in training or not (rows 3 and 4, F(1,18) = 0.88, p = 0.36). Consequently, the model performed as well on novel constructions with three embeddings as on trained such constructions. Because the level of accuracy was high, these results suggest that the model is capable of productively combining basic structures into novel multi-clause constructions. A concrete example of a construction that the model had not experienced in learning but produced correctly is instantiated by the sentence:

(8) `a woman that a man that the girl that is give -ing a beer to a dog jump -s with hit -s is carry -par by the mother .`

The combination of a transitive passive main clause with subject-relativized prepositional dative, oblique and intransitive embeddings was novel to the model, but it managed to correctly produce this combination of clauses based on its experience of simpler structures in different relative clause constructions. This accomplishment is quite remarkable given that sentence (8), which contains three center-embedded relative clauses, is particularly difficult to process for humans.

In previous experiments, the Dual-path model received sparse input in terms of sentence tokens, but it was exposed to most construction types in the target language and generalized syntactic knowledge to virtually all sentence tokens of these types. In the condition described above, the model received sparse input in terms of admissible construction types. I argued that the model is capable of generalizing familiar simple clause units to novel combinations in construction types not experienced during learning. Such structural generalization is possible because the model learns to correctly produce single-clause constructions and it abstracts a relativization principle based on simpler relative clause structures. This syntactic knowledge is then used to correctly produce novel relative clause constructions such as (8), which were not attested in the linguistic environment. The model achieved this feat by a data-driven, domain-general procedure. Hence, this result suggests that the ability for structural productivity in humans might be learnable from positive input, and that no language-specific biological mechanism need to be posited to explain this aspect of human linguistic behavior.

6.5 Lexical generalization

If there is one sense in which natural language is potentially infinite then in terms of its lexical openness. Novel words and expressions can be introduced into a language to coin idioms, or to denote novel entities, actions or events, and this happens every day. Humans have the ability to rapidly incorporate these words into their active knowledge of language. Once I have informed you, for instance, that a *klikusch* is a small, furry herbivore living in the woods you are able to understand a whole range of facts about this creature. In language acquisition, children learn the meaning of words in specific contexts and are subsequently able to use these words in novel contexts. Children may learn, e.g., that cats sleep most of the time, that cats play in the garden, and that cats sometimes chase dogs. Grasping the concept of a cat enables children to also understand that cats can be played with and that dogs sometimes chase cats. Familiar concepts can be comprehended in novel semantic contexts and familiar words can be used in novel syntactic contexts. This coherence of human thought and language has been labelled *systematicity*. Systematicity has been identified as a prime *explanandum* for models of language processing in general, and as a fundamental problem for connectionist models in particular (Fodor and Pylyshyn, 1988). Claims about systematicity are often phrased in terms of vague conditionals which are difficult to translate into meaningful empirical experiments. When framed in terms of learning, however, systematicity can be explicated more precisely as a form of lexical generalization (Hadley, 1994, 2004). According to Hadley a language processor displays

(i) *weak systematicity* if after learning it can process novel combinations of familiar words in familiar syntactic positions,

(ii) *strong systematicity* if (i) and it can process novel simple and embedded sentences containing familiar words in novel syntactic positions,

(iii) *strong semantic systematicity* if (ii) and words tested for property (ii) occupy novel thematic roles.[5]

In this section I adopt Hadley's explication and show that the Dual-path model displays these types of systematicity to a high degree. Based on the results from the previous section, I argue that the model satisfies an even stronger condition than (iii), namely *super-strong semantic systematicity.*

Weak systematicity was already demonstrated in Chapter 4, where the model was exposed to a small fraction of sentence tokens with at most one embedding admissible in the language. After learning, the model generalized to the complete language when tested on novel sentences composed of familiar constituents. Since strong systematicity is implied by strong semantic systematicity, I only tested the latter property in the model. To do this, the model was trained on the same distribution as in the previous section on structural generalization. In the training set, however, the word `cat` could occur only in the subject position and Agent role of simple-clause active transitive sentences such as `the cat chase -ed the dog.` The model was then tested on six different sets of 100 sentences each. The first test set contained sentences randomly drawn from the language which had one relative clause. This clause was a prepositional dative in which the indirect object or recipient slot could be occupied by any lexical item denoting an animate entity, except the word `cat`. The second test set contained sentences which had two relative clauses. In these sentences, the deepest embedding was a prepositional dative with the same lexical constraints as in the first test set. The third test set contained sentences which had three relative clauses, and likewise had such a prepositional dative as its deepest embedding. The remaining three test sets contained exactly the same sentences as the first three test sets, with the exception that the recipient of the dative action in the deepest embedding was always filled with the word `cat`. Thus, the model experienced the word `cat` in learning only in a one specific syntactic and semantic role in one specific type of single-clause utterance. The question was whether it could generalize to correctly using this word in a novel syntactic and semantic role in relative clauses of varying depth. Schematically, the experimental set-up is depicted in Figure 6.4. If the model can accomplish this lexical generalization, this would indicate that it behaves systematically in the sense of definition (iii). By having two matched test sets for each level of embedding, with only one critical lexical difference, we can measure the model's degree of systematicity as a function of sentence complexity. The results of testing strong semantic systematicity are shown in Figure 6.5 (page 172). As

[5] Hadley's formulation of strong semantic systematicity demands the assignment of "appropriate meanings to all words occurring in novel test sentences which (would or could) demonstrate the strong systematicity of the [processor]" (Hadley, 2004, p. 149). My rendition (iii) is stronger in that semantic content assigned to words is required to extend beyond linguistic experience too.

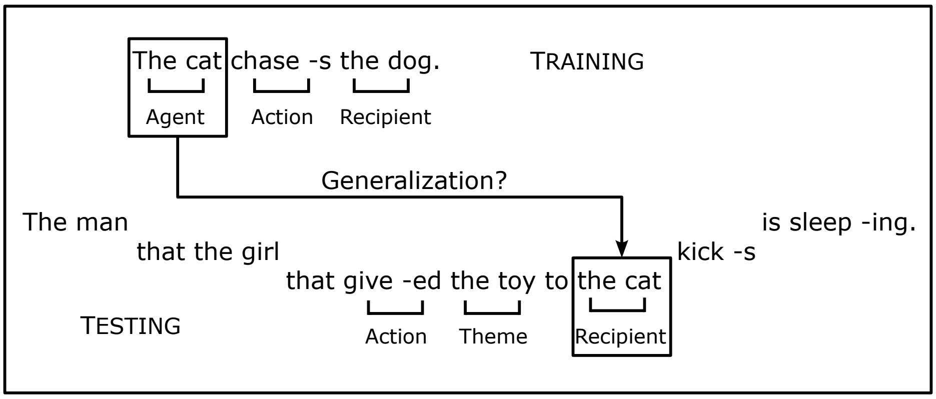


Figure 6.4: The constituent `cat` is trained in the AGENT role of non-embedded sentences only and the model is tested on sentences with `cat` in the RECIPIENT role of prepositional dative embeddings.

in the previous experiment, the model's production accuracy degraded with the number of embeddings in the test items. Sentences with one embedding were learned to perfection, sentences with two embeddings to >90%, and sentences with three embeddings to >65%. Pairs of trajectories represent the model's performance on two matched test sets. One trajectory in each pair shows the accuracy on sentences with the word `cat` in the recipient slot of the deepest dative embedding, the other shows the accuracy on the same sentences but with different words in this slot. It can be observed that for each level of embedding the model's behavior is very similar for both sets of test items, with small differences of less than 10% at epoch 200.000. The model's performance on the three sets of 'no cat recipient' utterances can be interpreted as baseline behavior for structures in which all lexical constituents have been observed in training in all syntactic and semantic roles. The difference in performance between the three sets of 'cat recipient' utterances and baseline indicates the degree of generalization of familiar lexical items to novel syntactic and semantic roles. Since these trajectories did not differ substantially the model displayed a high degree of strong semantic systematicity.[6]

The Dual-path model generalized familiar constituents (`cat`) from experience in simple clause structures to novel syntactic positions (indirect object) and novel thematic roles (recipient) in novel sentences at novel levels of embedding. Moreover, as in the previous experiment on structural generalization, a substantial proportion of the tested sentences were instantiations of constructions that the model did not experience in training (82.5% of the tested utterances). Hence, the model also performed lexical generalization inside structural generalization. I call this property

[6] A minor caveat should be mentioned here. Hadley demands that for strong systematicity "a significant fraction of the vocabulary of the training corpus must be presented in these novel positions" (Hadley, 2004, p. 149), whereas I only tested one lexical item. There is, however, no reason to assume that the model could not accommodate more novelty.

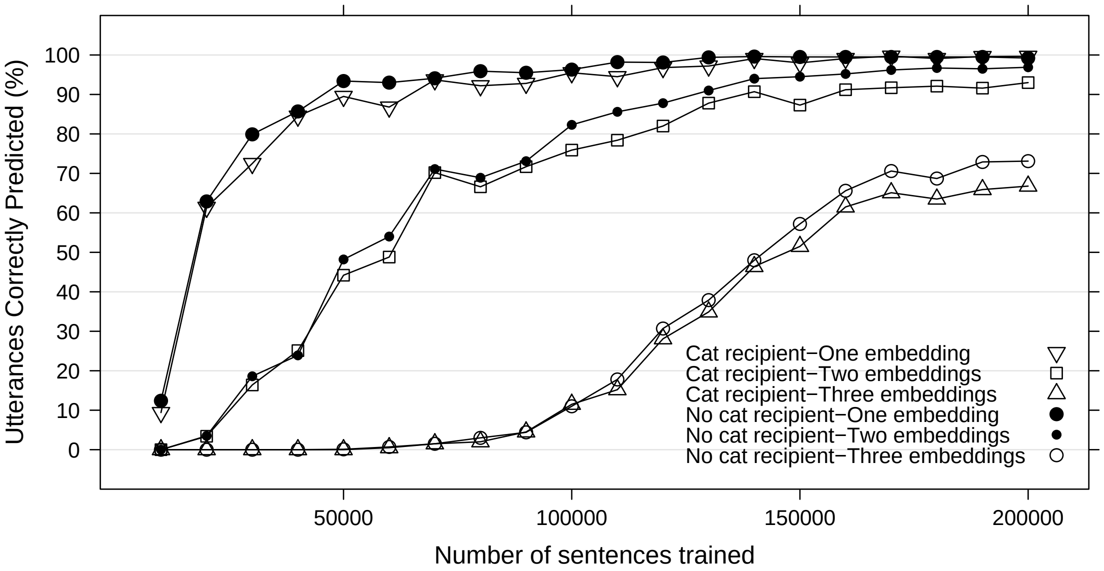


Figure 6.5: Strong semantic systematicity in the Dual-path model.

(iv) *super-strong semantic systematicity* if (iii) and words tested for property (iii) occur in novel constructions.

The Dual-path model can accomplish lexical generalization in the sense of (iv) for two reasons in conjunction. First, the model acquires syntactic knowledge in the form of argument structure constructions such as Transitive action or Dative transfer. These abstract frames determine the properties of verb arguments in each position, e.g., a dative recipient is always an animate noun. If a concept is understood by the model in that the mapping between concept and its word form is learned from examples, and if that concept has the right properties, it can enter into an argument slot of such a syntactic frame. And secondly, argument slots are controlled by the dynamic bindings in the what-where-system of the model. These bindings turn the thematic role nodes in the where-layer into semantic variables that can be instantiated by concepts not encountered in training. In the systematicity task, the model learned the dative frame from sample sentences not involving the word `cat`. Thus, it correctly activated the Recipient role in the where-layer after producing the preposition `to`. In the test utterances, this role was bound to the concept cat in the model's message input and activation spread along this binding. Because the model has learned the mapping from this concept to the word `cat` from simple-clause sentences in which `cat` occurred in subject position, it could activate `cat` in a novel semantic role, within a dative embedding of a novel construction. Hence, the combination of learned syntactic frames and dynamic bindings in the message allowed the model to go behave strongly systematic.

Although these bindings enabled lexical generalization, they did not guarantee systematicity. This is because the model also formed statistical expectations about the co-occurrence of words in the training corpus. In learning, the model records, for instance, that the preposition to is always followed by some nouns but not others (including cat). And it records that the noun cat is always followed by a transitive verb and never occurs in a dative frame. These expectations about transitional probabilities are formed in the model's sequencing system and this system competes for activation of lexical items with the message-lexical system at the output layer. Whether the model behaves systematically therefore depends on whether the information propagated via the dynamic bindings in the message-lexical system is sufficient to overwrite the statistical expectations of the sequencing system and win the competition for word activation (Figure 6.6). That the sequencing system occasionally wins this competition can be demonstrated by detailing the model's error profile in the lexical generalization task. I examined the model's production errors in the first 100 incorrect test sentences with three embeddings which had the word cat in the recipient slot of a prepositional dative in the deepest relative clause. Minor lexical errors, e.g., wrong articles, and errors in tense or aspect were ignored. The error profile is summarized in Table 6.5.[7] 37% of all errors occurred in the initial segment before the cat position. 21% of all errors, however, occurred at the cat position where the model produced a different noun instead (e.g., dog, man, etc.). This substitution error can be attributed to the expectations of the sequencing system about the class of lexical items which can occupy the recipient role of a dative frame. Furthermore, 33% of all errors occurred in the phrase immediately following the cat position, after cat was produced correctly. The output cat has been

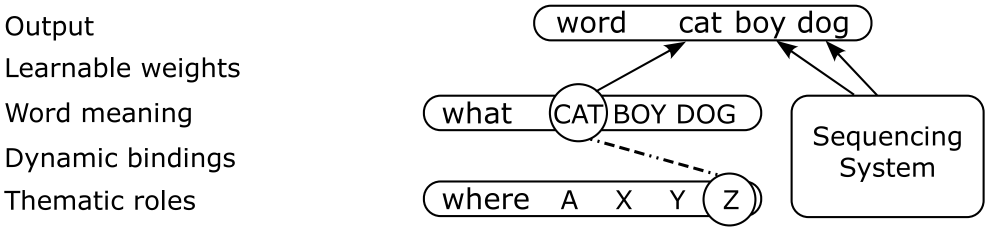


Figure 6.6: Dynamic bindings enforce systematicity and both pathways compete for word slots.

Before cat recipient	At cat recipient	Phrase immediately after cat recpient	Later in the sentence	Total
37	21	33	9	100

Table 6.5: Error profile for 'cat recipient' test items with three embeddings.

[7] Several utterances contained multiple errors as the model sometimes starts to produce gibberish once it made a severe grammatical mistake, such as incorrect attachment. I recorded the position of the sequentially first error in each sentence.

fed back to the input when this phrase is produced and triggers the model's statistical knowledge that `cat` should be succeeded by a transitive verb. Since the constituent `cat` completes the dative embedding, the following phrase always belongs to the clause in which the dative is embedded. Hence, these errors indicate that the model's lexical expectations (based on `cat` feedback) conflict with the syntactic structure of the complex utterance and this creates difficulty in resuming the superordinate clause. In sum, the majority of errors the model committed occur at or right after the dative recipient. This suggests that systematicity in the model is not a self-evident consequence of the functionality of the WHAT-WHERE-system, but a falsifiable empirical observation.

In the theoretical literature on systematicity, it is often argued that systematicity ought to be analytically derivable from the architecture of the processor and/or the nature of its syntactic representations (Fodor and Pylyshyn, 1988). Sometimes systematicity is even required to follow by nomological necessity (Fodor and McLaughlin, 1990). Neural networks usually do not satisfy these requirements due to their unstructured, distributed representations. Consequently, even if these systems were demonstrated to behave strongly systematic, they would not easily satisfy many critics of connectionism. In my view, it is difficult to justify that an empirical question about the nature of human syntactic representations should be turned into a philosophical dogma which excludes neural network architecture by fiat. The Dual-path model offers an interesting alternative in this debate. Systematicity does not follow analytically from the model's architecture nor its syntactic representations. Dynamic bindings between thematic roles and lexical meaning enable systematicity but whether this architectural feature is sufficient to overwrite statistical expectations of the learner is an empirical question. In this way, the Dual-path model approach might help to reconcile symbolic representations with the gradational nature of human syntactic processing.

6.6 Right-branching versus center-embedding

It was demonstrated that the Dual-path model generalized beyond its linguistic input in interesting ways. But how *natural* is this generalization behavior? This question is of critical importance if a cognitive model of syntactic development is to be considered relevant to psycholinguistic research. In light of the examples given in Section 6.3, it seems likely that humans can process complex sentences composed of novel combinations of familiar clause structures. Yet, this data is anecdotal at best. Human generalization capacities are difficult to determine experimentally because this would require recording the history of linguistic experience in individuals. One way to circumvent this problem would be to measure human processing accuracy on different complex structures and to estimate generalization capacities based on suitable corpus frequencies. To my knowledge, such studies have not been reported in the literature for the kind of structural generalization under discussion. I will therefore evaluate how natural the model's processing behavior is in terms of its differential learning and generalization of distinct relative clause constructions.

One dimension of distinction between relative clause constructions is their syntactic complexity. Right-branching constructions are considered less complex than center-embedded constructions because they can be described by a regular grammar whereas center-embedding requires a context-free grammar.[8] The relative difficulty of these two constructions in human sentence processing has been studied extensively in psycholinguistics. These studies unanimously found that center-embedded structures are harder to process than right-branching structures over a number of different measures such as accuracy in comprehension and recall or reaction times in grammaticality judgement (Blaubergs and Braine, 1974; Blumenthal and Boakes, 1967; Caplan et al., 1994; Foss and Cairns, 1970; Fodor and Garrett, 1967; Larkin and Burns, 1977; Marks, 1968; Miller and Isard, 1964; Stromswold et al., 1996). Bach et al. (1986) found the same differential behavior in German. Some studies also found that the degree of difficulty with center-embedded structures increased with the number of embeddings (Marks, 1968; Miller, 1962; Miller and Isard, 1964) and that center-embedded structures are facilitated by semantic constraints on noun-verb pairs in adults (Stolz, 1967) as well as children (Huang, 1983).

On the dominant view, this data can be explained by postulating performance constraints on human competence grammar in the form of working memory limitations (Berwick and Weinberg, 1984; Gibson, 1998; King and Just, 1991). In center-embedded, but not in right-branching structures, clauses are interrupted by intervening material. While processing these embeddings, the head nouns of center-embedded relative clauses must be held in working memory as unintegrated components of the matrix clause until the matching verb is encountered. In right-branching structures, on the other hand, every such noun-verb dependency is resolved before the processing of a new clause begins. Thus center-embedded structures are taxing working memory more than right-branching structures and this might explain human differential processing. Furthermore, if working memory degrades with the amount of intervening material this account might explain why human processing difficulty correlates with depth of embedding.

In this section I first investigate whether the Dual-path model behavior qualitatively matches human differential performance on these two types of structures. I will then assess the validity of the working memory hypothesis in the model by analyzing its error profile in sentence production. This simulation required extending the input language used previously. So far this language only allowed singular nouns and verb forms and the model did not have to maintain noun-verb agreement. Number agreement, however, creates long-distance dependencies when the language permits embeddings and hence errors in agreement can tap into working memory failure. I implemented number by adding a plural marker `-s`, a distinct simple present marker `-ss` and the plural auxiliaries `are` and `were` to the language. These markers were treated as separate lexical

[8] Strictly speaking, this is not true for string languages but it is true for natural language. For example, a formal grammar can be constructed which allows center-embedding and long-distance dependencies and generates the language $\{ab^nc \,|\, n \geq 1\}$ which is regular. If, however, the non-terminal S which self-embeds is rewritten by at least two different terminals to the left and right of S, the language becomes context-free, e.g., $S \rightarrow aSb$, $S \rightarrow ab$ which generates $\{a^nb^n \,|\, n \geq 1\}$.

items as exemplified by the sentence:

(9) `the dog -s that a brother jump -ss with were push -par by a boy .`

The model was trained on a set of 10.000 randomly generated sentences from this language which again was combinatorially complete for a maximum of three embeddings. The distribution of embedding depth in the input followed Figure 6.2 (page 167). For each level of embedding the amount of training on center-embedded and right-branching structures was balanced. Thus, the model experienced exactly the same number of single, double and triple embedded such structures in learning. In testing, the model was exposed to 100 right-branching and 100 center-embedded structures for each level of embedding. Test structures classified as right-branching or center-embedded, had to be genuinely right-branching or center-embedded. Sentences which had, e.g., a center-embedding inside a right-branching relative clause were not permitted in testing as a center-embedded structure. Such mixed structures, however, were allowed in training but were not counted as either structure when balancing the amount of training on each structure. Likewise, sentences in which a relative clause was attached to a dative object were not counted as either structure in training or testing. Because the matrix clause is disrupted by such relative clauses these structures are not genuinely right-branching and because one noun-verb dependency is resolved before the relative clause they do not qualify as center-embedded either. Table 6.6 (page 178) shows examples of the kind of structures which were tested and excluded from testing.

In sentences with three embeddings, notice that the deepest relative clause in both structures can be subject- or object-relativized but the first and second relative clause must be subject-relativized in right-branching structures and object-relativized in center-embedded structures. Evidence from many languages indicates that subject-relative clauses are easier to process than object-relative clauses, for children and adults. This processing difference at the clausal level might further exacerbate the processing of center-embedded structures in humans. Differential performance on subject- and object-relative clauses has also been explained in terms of working memory limitations (Gibson, 1998; King and Just, 1991; Wanner and Maratsos, 1978), and I will look at this contrast in more detail in Chapter 8.

When ten model subjects were trained and tested in the described way, the mean performance data in Figure 6.7 (page 177) was obtained.[9] This graph displays a number of interesting properties. First, the model's performance degrades with the depth of embedding. Sentences with one relative clause and sentences with two right-branching relative clauses were learned almost to perfection. Doubly center-embedded structures reached around 80%, triple right-branching structures around 65% and triple center-embedded structures around 35% sentence accuracy. Broadly speaking, this degradation is

[9] In this condition more training was required for the model to reach a level of accuracy comparable to previous experiments because the target language was more complex. It was also required to add Gaussian noise to the connection weights which feed into the HIDDEN-layer to elicit better generalization behavior. See Appendix A for details on this issue.

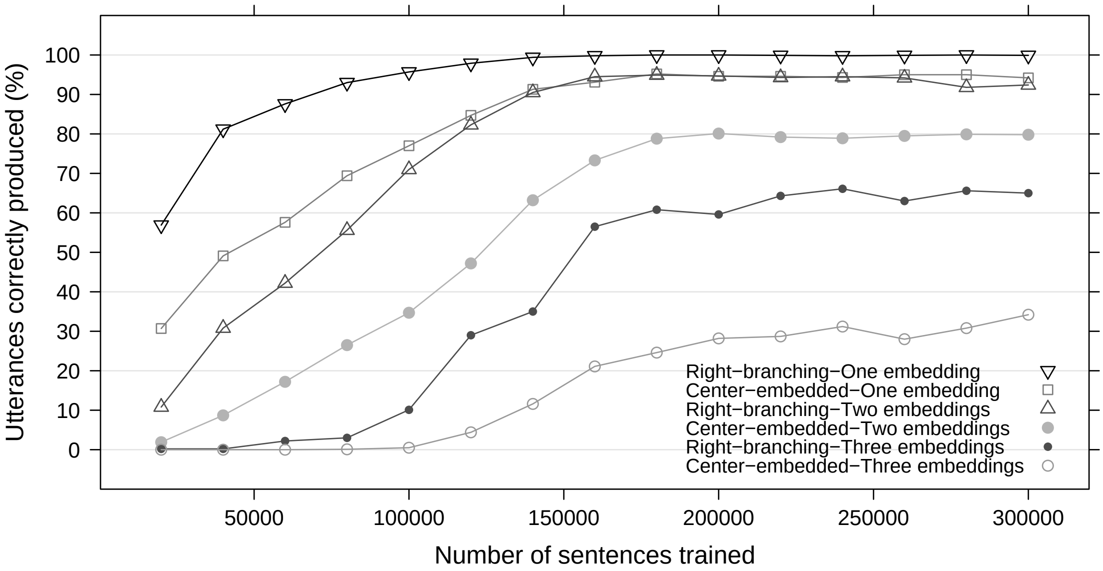


Figure 6.7: Comparison of sentence accuracy for right-branching and center-embedded structures by depth of embedding.

in line with human sentence processing. It falls out of the architecture of (and input to) the model and no external performance limitations need to be stipulated to obtain this behavior. The more complex a structure is the harder it is for the model to extract regularities from the surface form. Since complex structures were sparse in the input and the model learned all structures over one set of connection weights it was more difficult for the model to memorize syntactic knowledge of more complex structures. These factors—hampered extraction and memorization—explain why performance degrades with depth. Secondly, the developmental trajectories reveal a substantial difference in learnability between right-branching and center-embedded structures. For each level of embedding, right-branching structures were learned faster and to a higher degree of accuracy at the end state. In addition, performance on center-embedded structures rapidly degraded with depth. Qualitatively this behavior matches human performance attested in the psycholinguistic studies quoted above. And third, center-embedded structures of depth n developed in closer proximity to right-branching structures of depth $n+1$ than to right-branching structures of depth n. For instance, double right-branching structures developed almost like single center-embedded structures and reached the same level of accuracy. This indicates that sentence length and the number of embedded clauses were not the most important determinants of the model's performance.

In order to trace the cause of differential behavior, I inspected the errors the model committed in testing on center-embedded and right-branching structures with three embeddings. This analysis is also used to evaluate the hypothesis that working memory limitations are responsible for the differential processing of these structures. In total 100 erroneous productions were examined, the first 50 incorrect right-branching and cen-

Construction type	Example sentence	Mean length (# lex. items)
Right-branching, 1 relative clause	the man -s were hit -par by a woman that was push -par by the sister -s .	17.0
Center-embedded, 1 relative clause	a brother that is run -ing with the sister was show -ing a orange to the girl .	15.8
Right-branching, 2 relative clauses	a cat present -ss the toy to a brother that walk -ed with the father -s that the man give -ss a stick to .	24.7
Center-embedded, 2 relative clauses	the girl -s that a brother that was show -ing the pear to the boy was hit -par by were jump -ing .	23.7
Right-branching, 3 relative clauses	the woman -s were carry -ing the dog that jump -ed with the nurse -s that are kick -par by the sister -s that a mother show -ss a kite to .	32.0
Center-embedded, 3 relative clauses	the woman -s that the man -s that the nurse that a mother is show -ing the milk to kick -ss walk -ed with were hit -par by the sister .	31.5
Mixed, excluded from testing	the nurse -s give -ed the toy to the woman -s that a sister that was run -ing was walk -ing with .	n/a
Clause-final prepositional phrase, not in test set	the boy -s throw the ball that the brother that is sleep -ing give -ed the dog -s to the woman -s .	n/a

Table 6.6: Examples of tested and excluded structures in the center-embedding vs. right-branching experiment.

ter-embedded structures, respectively. Error classification was based on the sequentially first error in each incorrect utterance. A summary of error types and their frequencies is shown in Table 6.7.

Error profile for *right-branching* constructions

Clause	Verb	(Number)	Noun	Structural	Other	Subtotal
Main	0	(0)	1	0	0	1
First	3	(0)	2	6	0	11
Second	3	(1)	4	3	0	10
Third	5	(2)	3	12	8	28
Total	11	(3)	10	21	8	50

Error profile for *center-embedded* constructions

Clause	Verb	(Number)	Noun	Structural	Other	Subtotal
Main	3	(1)	0	4	0	7
First	8	(2)	0	2	0	10
Second	11	(6)	1	0	0	12
Third	5	(0)	2	12	2	21
Total	27	(9)	3	18	2	50

Table 6.7: Error classification for 100 incorrect right-branching and center-embedded structures. *Verb* errors comprised wrong tense, aspect, stem and number agreement. The amount of number agreement errors is given in parentheses. *Noun* errors comprised wrong articles, nouns and number. A *structural* error was assigned when active and passive voice were confused, a subject-relative clause was turned into an object-relative clause or vice versa. Errors were classified as *other* when a lexical omission, repetition, insertion or substitution occurred which could not be interpreted as one of the other error types. Attachment errors did not occur in any inspected utterance.

Recall that the language used to train the model had noun-verb number agreement. Together with center-embeddings this created long-distance dependencies. After a center-embedding is completed, clauses are resumed with verbs or auxiliaries which have to agree in number with the head noun of the interrupted clause. If limitations on working memory made center-embedded structures particularly difficult to process, we should observe many production errors at these sentence positions. Thus, we should expect

(i) head noun/matrix clause verb number agreement errors to occur frequently

in center-embedded structures. Moreover, we should observe

(ii) more such errors the larger the linear distance between the head noun and its verb is.

In other words, the working memory hypothesis predicts that in center-embedded structures more agreement errors should occur at the main clause verb/auxiliary position than at verb/auxiliary positions in embedded clauses. In right-branching structures on the other hand all noun/verb agreement is resolved within each clause before the next embedding starts. Hence no clause is taxing working memory more than any other. Thus we should observe that

(iii) agreement errors at verb/auxiliary positions are uniformly distributed over all clauses

in right-branching structures. Secondly, we should find that

(iv) for each clause level (except the third embedding) such errors are less frequent in right-branching than in center-embedded structures

if working memory limitations cause the observed processing differences between the two structures.

The error profiles for these two structures in Table 6.7 reveal that center-embedded structures caused more than twice as many verb errors than right-branching structures and three times as many agreement errors.[10] Verb (agreement) errors account for 22% (6%) of all errors in right-branching structures and for 54% (18%) in center-embedded structures. This general picture seems to confirm the working memory hypothesis. Furthermore, agreement errors in the main clause, the first and second embedding are more frequent in in center-embedded structures than in right-branching structures. Thus, prediction (iv) is supported by the data. When we look at predictions (i)–(iii), however, the working memory hypothesis is not so well-supported. Both the number of verb and agreement errors increased with clause depth in right-branching structures and no such errors occurred in the main clause which is not in line with prediction (iii). This fact can be explained in terms of the close similarity of right-branching main clauses and simple-clause structures. Main clauses of right-branching structures—like simple-clause structures—are uninterrupted by an embedding and overtly express all event participants, in contrast to embeddings in which one participant is gapped. Thus right-branching main clauses benefit from exposure to simple-clause structures and these structures are very frequent in the input environment.

Although verb errors occurred more frequently in center-embedded than in right-branching structures, the proportion of agreement errors out of all center-embedding errors is still rather low (18%). In addition, the proportion of agreement errors out of all verb errors is roughly the same for right-branching and center-embedded structures (around 30%). Consequently, there is only weak support for prediction (i) in the data. Moreover, prediction (ii) which most explicitly reflects the working memory hypothesis is disconfirmed by the error profile. Agreement as well as verb errors occurred least frequently in main clauses and less frequently in the first embedding than in the second.

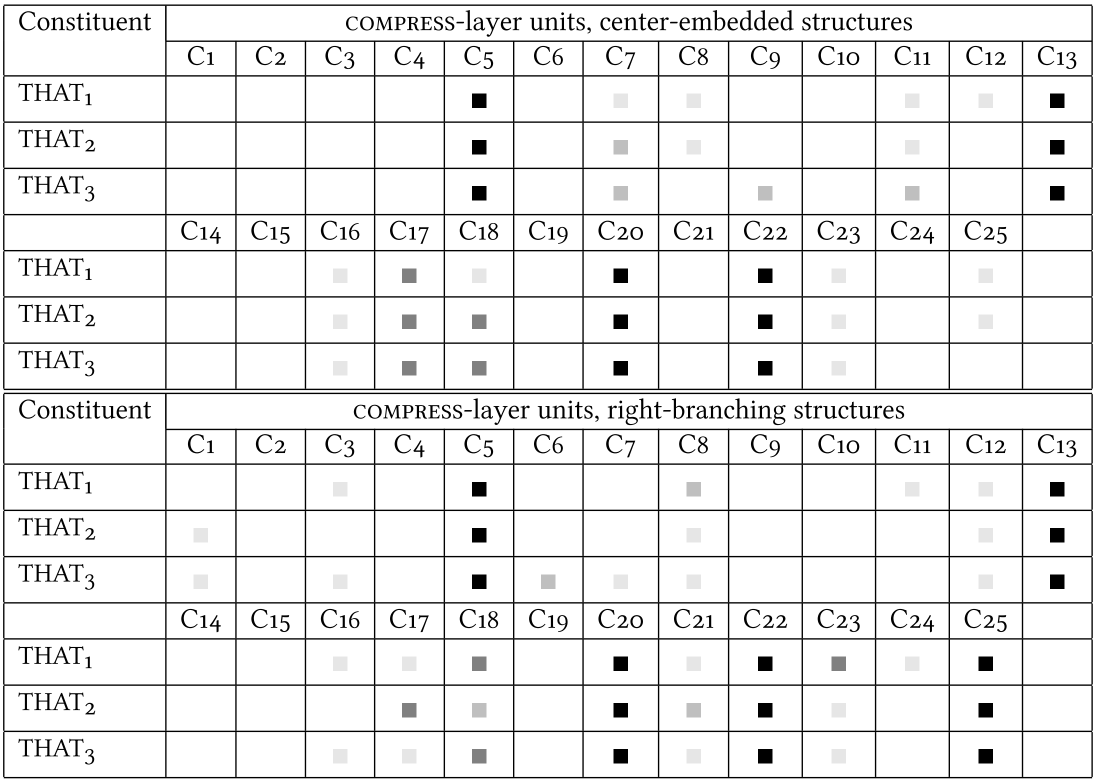
[10] Neither difference, however, was statistically significant across clauses; verb errors $t(3) = 1.04$, $p = 0.375$, number errors $t(3) = 2.38$, $p = 0.097$.

This behavior undermines prediction (ii) for center-embeddings. Hence, the Dual-path model does not support the claim that long-distance dependencies cause processing difficulties due to working memory limitations.

The model's differential performance can be explained in a multifactorial way by a combination of similarity and frequency considerations. The distribution of errors over clauses is very similar for both structures, the least amount of errors occurred in the main clause, the most in the third relative clause. The most conspicuous difference in the error profiles of both structures lies in the amount and distribution of verb errors. As remarked above, right-branching main clauses are similar to single-clause utterances which are frequent in the input. Therefore no verb errors occurred in these clauses. Center-embedded main clauses are interrupted by embeddings of various length and depth and overtly express every event participant from the message. Thus, they are structurally different from clauses inside center-embeddings. But because every center-embedding in the training language has a main clause, these structures are more frequent than, e.g., second-level center-embedded clauses. This might explain why we observe relatively few verb errors in the main clause despite long and nested embeddings disrupting the noun/verb dependency.

The deeper the level of center-embedding, the less frequent these clauses (and the structures to which they belong) are in training. This leads to sequences of word categories which are increasingly sparse in the model's input. In case of a triple center-embedded structure, e.g., sequences of word categories such as NOUN VERB$_3$ VERB$_2$ VERB$_1$ VERB$_0$ PER can occur where verb subscripts denote clause level and PER is the end-of-sentence marker. These sequences are specific to triple center-embedded structures, i.e., they do not occur in any other structure. Because these structures are sparse, such subsequences are sparse and this causes processing difficulty in the model. In right-branching structures, on the other hand, no such unique subsequences occur. Here, verbs are never followed by verbs but always by object nouns or the end-of-sentence marker. These subsequences are vastly more frequent in the input because they are shared with single-clause utterances. This difference in substructure similarity and frequency between right-branching and center-embedded structures can explain differential processing at all levels of embedding. Substructure frequency can also explain the distribution of verb errors for center-embedded structures across clauses. Substructures such as VERB$_1$ VERB$_0$ PER can occur in every center-embedded structure regardless of its depth. VERB$_2$ VERB$_1$ VERB$_0$ PER, however, occurs only in double and triple embedded structures, VERB$_3$ VERB$_2$ VERB$_1$ VERB$_0$ PER only in the latter. That the second embedding is resumed with a verb after interruption by the third embedding is only ever witnessed by the model in triple center-embedded structures. Consequently, we observe most errors at the verb position of the second embedding (total: 11, Table 6.7, page 179). That the first embedding is resumed with a verb after interruption by the second embedding is witnessed by the model in double and triple center-embedded structures, hence this pattern is more frequent and we observe less errors at this verb position (total: 8, Table 6.7, page 179). Similarly, it can be argued why the lowest verb error rate occurred in main clauses of center-embeddings. And finally, all clauses in right-branching structures are similar to each other and to single-clause utterances in that they are non-interrupted, whereas clauses in center-embedded structures are interrupted by embeddings of different length. Hence in learning the model gains more experience with all clause types in right-branching structures than with any clause type in center-embedded structures which facilitates better overall performance on right-branching structures at each level of embedding.

Although the Dual-path model, which is a statistical learning mechanism, is sensitive to substructure frequencies, it is not processing complex sentences as linear strings of words. The model is also sensitive to the distinct hierarchical structure of these two constructions. Evidence for this property can be obtained by visualizing the internal representations developed by the model during syntactic development. The model represents abstract structural knowledge in its sequencing pathway. To find representational differences between right-branching and center-embedded structures the compress-layer is therefore a natural model component to look at. Similar to the procedure described in Section 5.2.6, I recorded the activation states of the compress-layer while the model was correctly producing ten sentence samples with three relative clauses of each type at the end of training. These state vectors were then averaged componentwise and quantized into five discrete activation levels. Figure 6.8 displays the outcome of this procedure for the three pronoun positions in the two utterance types. The top



Constituent	compress-layer units, center-embedded structures												
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
THAT$_1$					■		░	░			░	░	■
THAT$_2$					■		▒	░			░		■
THAT$_3$					■		▒		▒		▒		■
	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	
THAT$_1$			░	▓	░		■		■	░		░	
THAT$_2$			░	▓	▓		■		■	░		░	
THAT$_3$			░	▓	▓		■		■	░			

Constituent	compress-layer units, right-branching structures												
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
THAT$_1$			░		■			▒			░	░	■
THAT$_2$	░				■			░				░	■
THAT$_3$	░		░		■	▒	░	░				░	■
	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	
THAT$_1$			░	░	▓		■	░	■	▓	░	■	
THAT$_2$				▓	▒		■	▒	■	░		■	
THAT$_3$			░	░	▓		■	░	■	░		■	

Figure 6.8: Comparison of activation states of the compress-layer for center-embedded and right-branching structures at the pronoun positions.

half of the table shows activation states for center-embedded and the bottom half for right-branching structures. It can be observed that units C5, C13, C20 and C22 are fully active in both structures. Of these four units C5 and C22 are highly specialized units in that they were only active at pronoun positions, whereas C13 and C20 showed some activation at other sentence positions as well, in particular verbs.[11] We see some minor differences in the activation levels of units C17, C18 and C23 between these two structures. The largest difference, however, occurred at unit C25. While this unit is virtually silent when processing center-embedded sentences, it is fully switched on at all pronoun positions when producing right-branching structures. In particular, this unit is only ever active for pronouns but no other constituents and this pattern consistently held for right-branching structures with one, two and three relative clauses.[12] Since `that` is a purely functional constituent which introduces embedded clauses, unit C25 is fully specialized in marking the difference between center-embedded and right-branching relative clause attachment. Hence, the Dual-path model was representing the distinct hierarchical organization of the two structures in its sequencing pathway.

Standard explanations of the contrast between right-branching versus center-embedding in human processing often invoke syntactic complexity and performance constraints such as working memory limitations. An alternative hypothesis would be that syntactic structures which are relatively complex, such as center-embedded sentences, are more difficult to learn, store and produce than relatively simple structures, such as right-branching sentences. One reason for this difficulty might be that center-embedded structures are less similar at the clausal level to single-clause structures, which are very frequent in the input, than right-branching structures. Another reason might be that center-embedded sentences contain substructures which are not shared with other input structures whereas right-branching sentences do not. Such asymmetries in similarity might differentially facilitate (or hamper) the grammatical encoding of these two structures during learning and lead to the observed differences in production. In this way, the input distribution of a statistical learner might be skewed towards one structure and against another despite the fact that both structures themselves occur with the same frequency (as in the described experiment). The presented results can be interpreted as supporting this hypothesis. In syntactic development, right-branching sentences profit more from exposure to other input structures than center-embedded sentences. Therefore they are acquired, memorized and activated more easily than center-embedded structures. This frequency-based explanation suggests that it might be unnecessary to stipulate performance constraints on competence grammar external to the language system—such as working memory limitations—in order to account for the differential processing of right-branching and center-embedded structures in humans.

[11] This is not visible in Figure 6.8 but only in the full COMPRESS-layer pattern for complete sentences. These are not reproduced here for the sake of brevity.

[12] Again, this is not observable in Figure 6.8 but was verified in the full COMPRESS-layer graphs for right-branching structures of various depth.

6.7 Recursive productivity

In this section I will review some recent arguments by Hauser et al. (2002) which suggest that recursion in natural language is a biologically endowed category unique to human communication. I will then sketch what in my view is the right way of looking at recursion in language processing and show that the Dual-path model can explain recursive productivity satisfactorily.

6.7.1 Is recursion uniquely human?

Mathematical recursion is a procedure to define objects (e.g., functions, sets) by reference to themselves. In natural language, recursion allows the embedding of one phrase, clause or sentence within itself. According to Hauser et al. (2002), recursion is a core property of the human language faculty. They argued that recursion is the only mechanism that both distinguishes language from other human cognitive capacities and separates human language from animal communication. In other words, except for recursion everything about language is either not uniquely human or uniquely human but not specific to language. These strong claims about recursion have been criticized from many different angles. For instance, Pinker and Jackendoff (2005) have argued that other aspects of language are not recursive but uniquely human and language-specific, such as phonology and morphology.[13] Others have questioned the evidential basis for claiming that recursion is uniquely human. In support of their position Hauser, Chomsky and Fitch adduce evidence from experiments on artificial grammar learning in non-human primates, e.g., their own study (Fitch and Hauser, 2004) which ostensibly shows that cotton-top tamarin monkeys are not capable of recursion. In these experiments it was argued that tamarins could learn membership for stringsets such as $\{abab, ababab\}$ but not stringsets such as $\{aabb, aaabbb\}$, whereas humans could learn both. The former language was considered representative for being generated by a regular grammar, the latter representative of a context-free grammar with center-embedding.[14] On the other hand, Gentner et al. (2006) have recently argued that even starlings could learn the $\{aabb, aaabbb\}$ stringset, undermining the case for the uniqueness of human recursion. But of course these are finite sets and syntactic complexity separates the two sets only for infinitary extensions of these languages. Hence, it is questionable whether mere discrimination of sets is sufficient to claim that the capacity for recursion has been detected. One would in addition have to exclude the possibility that the cognitive procedure to determine the grammaticality of strings is non-recursive. Corballis (2007), for example, argues that the starlings of Gentner et al. (2006) could have accomplished this task by a simple counting strategy without any knowledge of the dependencies indicated by subscripts in $\{a_1a_2b_2b_1, a_1a_2a_3b_3b_2b_1\}$. The experiments which lead Fitch and Hauser (2004) to claim that humans, in contrast to tamarins, could learn a context-free

[13] This criticism was followed by a rejoinder from Fitch et al. (2005) and a subsequent reply by Jackendoff and Pinker (2005).

[14] Cf. Pullum and Rogers (2006) for a criticism of this assumption.

grammar were methodologically flawed in similarly fundamental ways as was pointed out by Perruchet and Rey (2005). They attempted to replicate the Fitch and Hauser data for humans but in addition tested their subjects on strings which violated syllabic dependencies in center-embeddings. They found that subjects were able to learn the $\{aabb, aaabbb\}$ stringset but only based on acoustic cues, they displayed no sensitivity whatsoever to the underlying grammar.[15] Perruchet and Rey concluded that their data were "consistent with the hypothesis that human participants performed the test as a simple perceptual discrimination task" (p. 310). Needless to say, this issue requires further study, but if this is true, not even humans draw on recursive processing to determine the grammaticality of word sequences (in the framework of artificial grammar learning). As a consequence, virtually every central claim made in Hauser et al. (2002) is rendered either false or vacuous.

One might ask, then, what the epistemic status of recursion in language processing might be. Rather than being an empirical fact, recursion has been utilized predominantly as a conceptual tool to explain other aspects language such as productivity and infinity. This is most obvious in the Hauser et al. (2002) paper where the authors state that language has the "capacity for limitless expressive power, captured by the notion of discrete infinity" (p. 1576) and that humans have the "capacity to recombine meaningful units into an unlimited variety of larger structures, each differing systematically in meaning" (ibid.). These properties of productivity and infinity are taken as self-evident[16] and recursion is offered as an explanatory notion:

> FLN takes a finite set of elements and yields a potentially infinite array of discrete expressions. This capacity of FLN yields discrete infinity (a property that also characterizes the natural numbers). (ibid., p. 1571)[17]

The relation between infinity, productivity and recursion, however, is not that straightforward and simplistic. Recursion is neither necessary nor sufficient to explain either infinity or productivity. It is well-known, for instance, that every primitive recursive function can be translated into a function which is defined by pure iteration without recursive calls.[18] Hence recursion is not necessary to generate infinite sets. Moreover, the mechanism of recursion in language does not guarantee that an infinite set is generated. A context-sensitive grammar with non-trivially recursive rules which generates only one string can easily be constructed.[19] Unless recursion is assumed to never terminate—an assumption of infinity by itself—recursive rules in syntax do not yield

[15] That is, subjects could reliably classify strings based on acoustic patterns from the training phase, but did not detect violations in syllabic dependencies, and there was no interaction between the two.

[16] "The core property of discrete infinity is intuitively familiar to every language user" (Hauser et al., 2002, p. 1571).

[17] FLN is Hauser, Chomsky and Fitch's acronym for the 'faculty of language in the narrow sense' which, on their view, comprises but recursion.

[18] Cf. Odifreddi (1989).

[19] Cf. Pullum and Scholz (2008). The recursive grammar $\{S \rightarrow AB, B \rightarrow BB, A \rightarrow a, B \rightarrow b/a\_\_, B \rightarrow c/ab\_\_\}$ generates the single string abc.

linguistic infinity. Thus, infinity is not derivable from recursion, every argument to this end is circular. From the point of view of linguistic theory there might be good reasons to study human language in infinitary models, especially when one is concerned with grammatical competence within the generative tradition. It must be clear, though, that infinity is a modelling choice, not a consequence of recursion which itself is not an empirically demonstrated property of human language processing.

6.7.2 Recursion and productivity

In contrast to infinity, linguistic productivity is a rather uncontroversial notion. Undoubtedly, humans have the capacity to produce and understand novel utterances they have not experienced in communication before.[20] Hauser, Chomsky and Fitch seem to assume that productivity and 'discrete infinity' are interchangeable. A productive language processor can create an infinity of utterances from finite building blocks, and if it can do so, it can be called productive. To belabor the obvious, both entailments are disputable. The game of chess has an estimated upper bound on state-space complexity of 5×10^{52} and a game-tree complexity of 10^{123}, both large but finite numbers (Allis, 1994). It is quite unlikely that a player of chess or in fact any two human beings will ever play the same game twice. Although the 'grammar' of chess is highly constrained, creatively productive game play is not ruled out by a finite number of board configurations. Conversely, formal mechanisms for generative infinity need not qualify as productive. As Pullum and Scholz (2008) pointed out, recursive rules such as 'Adjective → very Adjective' may generate an unbounded number of phrases 'very nice', 'very, . . ., very nice', etc., but these would hardly be considered the core of linguistic productivity (let alone creativity).[21] For these reasons infinity is not interchangeable with productivity and furthermore recursion is neither necessary nor sufficient for explaining productivity.

Attempting to account for productive infinity by means of recursion is misguided in at least two ways. Generative infinity is a modelling assumption which cannot be justified through recursion and linguistic productivity in humans is not explained by recursion alone. Since there is no explanatory relationship between recursion and productivity, recursion in language needs to be motivated independently. Two aspects of recursion must be distinguished, recursion in syntactic modelling and recursion in human language processing.

Some of the appeal recursive principles have to theoretical linguists might lie in their high level of abstraction and descriptive parsimony in the representation of natural language syntax. The syntax of embedding one clause within another, for instance, can be described by simple and general recursive rules. Alternatively, it can be described as the recombination of clausal construction types by means of relativization. But this would require listing all syntactically admissible combinations of non-embedded constructions.

[20] Ideally, every PhD thesis consists of novel utterances, modulo quotations, produced by the doctoral candidate and intelligible to the dissertation committee.

[21] One is tempted to label the application of such recursive rules to create novel linguistic utterances 'moronicity' rather than productivity.

There is, however, a price to be paid for abstraction and parsimony, which is potential overgeneration and empirical inadequacy. Based on corpus analysis Verhagen (2008), argued that syntactic constructions which look like paradigmatic cases of abstract recursion at the clausal level (e.g., causative constructions and long-distance *wh*-movement) may in fact be instances of concatenating lexically specific templates. No general rule for embedding one clause within another is necessary (or even adequate) to model the syntax of such constructions, and the usage of such rules in processing is not licensed by the corpus data.

The psychological reality of recursion in language processing has not been demonstrated in the literature. Results such as Perruchet and Rey (2005) rather seem to point in the opposite direction.[22] A motive for adhering to recursion in language processing might be that recursion has been identified as an important organizational principle in cognitive domains other than language, such as hierarchical decomposition in planning, navigation, problem-solving, or goal-directed action, and hierarchical composition in grouping, object combination, or tool use. If it can be demonstrated that recursive behavior is operant in other cognitive domains and recursion can be established as a useful or even indispensable principle for describing natural language syntax, it is a small step to asserting the psychological reality of recursion in language processing. Greenfield (1991), for instance, has suggested that the human capacities for hierarchical language production and manual action are functionally analogous and argued for an evolutionary homologue (confer Arbib and Rizzolatti, 1997 and Steedman (2002) for similar views). This position differs from Hauser, Chomsky and Fitch's view that recursion, although exapted from computational mechanisms outside the domain of language (e.g., 'number, navigation, and social relations'), is language-specific and did not evolve by homology (neither within the human species nor from a common ancestor). It also differs from Pinker and Jackendoff's view that language in general is a unique adaptation for communication (Pinker and Jackendoff, 2005). It must be pointed out, however, that functional analogy does not establish an evolutionary connection between cognitive domains unless strong neuroanatomical evidence is provided. More importantly, functional analogy does not preclude the possibility that, say, motor planning and syntactic processing are realized by computationally distinct mechanisms.

6.7.3 The innateness of recursion

Hauser et al. (2002) devote large parts of their paper to describing a variety of different hypotheses about the evolutionary origin of the language faculty broadly conceived, and recursion in particular. But how did they arrive at the fundamental conviction that recursion must be part of our biological endowment in the first place? To recap, they first introduced the notion of 'discrete infinity' to which language users are 'intuitively

[22] These results, however, are very limited in nature because they derive from artificial grammar learning where nonadjacent dependencies have no semantic value. As the authors suggest themselves, the study of recursion across species using string-languages from the Chomsky hierarchy may be a 'conceptual dead-end'.

familiar'. They invoked the concept of recursion as an *explanans*, without offering any clarification what exactly they mean by recursion and how recursion is generating 'discrete infinity'. I argued above that in any case this line of reasoning was circular. Then they suggest the non-learnability of natural language syntax based on the observation that children only ever experience a finite amount of linguistic input:

> [T]here are in principle infinitely many target systems [...] consistent with the data of experience, and unless the search space and acquisition mechanisms are constrained, selection among them is impossible. (p. 1577)

In light of the fact that they quote results by Gold (1967) in support of this claim, this is quite a peculiar statement since Gold, among other things, has proven the learnability of infinite languages from finite data. But suppose we grant this point and even accept the conclusion that these constraints must be innate (which they do not argue for), then it still remains open why these are constraints on the faculty of language in the narrow sense (as they claim on page 1577) and not domain-general constraints on human learning mechanisms (as they suggest in the quotation above). Suppose furthermore that this leap in argumentation can be justified, then these innate constraints are constraints on recursion (since FLN is co-extensive with recursion according to Hauser, Chomsky and Fitch). It is not clear why this should *ipso facto* render recursion itself an innate mechanism which is susceptible to biological evolution. In other words, even if we buy into every single assumption that Hauser et al. (2002) put forth to establish recursion as a core property of the human language faculty, their argumentation does not validate the idea that recursion should be conceived of as an innate principle of language processing rather than an acquired capacity. It is a plausible alternative worth investigating whether productive linguistic behavior (which can be *described* as recursive processing) is learnable through innately constrained, domain-general mechanisms, based on linguistic experience. I will now sketch such an alternative approach and present some evidence that this could be accomplished by a data-driven learner.

6.7.4 A proper *explanandum*

The powerful computational mechanism of non-terminating recursion is often postulated in theoretical linguistics to account for infinite productivity. This capacity is attributed to human linguistic competence. In actual linguistic performance the ability to produce or comprehend recursive embeddings rapidly degrades with depth (see Section 6.7). Christiansen (1992) argued that the distinction between infinite recursive capacities and observable linguistic behavior should be abandoned. His arguments against the competence/performance distinction are based on the architecture of neural network models. In these models the physical location of stored syntactic knowledge and the language processor itself are inseparable. Hence there is no knowledge base of linguistic competence which, when restricted appropriately by external constraints, yields observable performance; the distinction collapses for architectural reasons. It could also be argued that the competence/performance distinction should be rejected on methodological grounds. What makes the distinction problematic is that competence attributes idealized knowledge to the language system which has no measurable or observable consequences. By definition, competence abstracts away from performance constraints such as limited memory, time and attention. Thus, competence puts Turing's ghost into the biomechanics of our resource-constrained brains. All 'articulations' of competence are filtered through the deficiencies of our production-comprehension system. Hence, any incompatibilities between linguistic behavior and the predictions of competence grammar can always be attributed to the failure of the procedures which access and utilize linguistic knowledge to produce and comprehend utterances. Non-terminating recursion is a theoretical entity which lives in competence grammar, and since it has no observable consequences, it is not a particularly useful assumption. For this reason, it is not an interesting *explanandum* for any theory of natural language processing, connectionist or not.

Eliminating non-terminating recursion as an *explanandum* does not obviate the need for an explanation of bounded recursive productivity in humans. In my view, the following aspects of recursive productivity need to be accounted for by any viable model or theory of language processing:

(i) How does the human language system recombine familiar clausal constructions into novel, hierarchically structured utterances not experienced during acquisition?

(ii) More specifically, how can (i) be achieved for utterance types with more levels of embedding than encountered in learning?

(iii) Why do productive capacities such as (ii) degrade with the depth of embedding in human performance?

(iv) Can the combination of (ii) and (iii), describable as *gracefully terminating recursive productivity*, be achieved without learning explicit rules of recursive composition, or does it require such rules plus external constraints?

In Section 6.3 above I argued that the productive capacity (i) can be explained in the Dual-path model for constructions with three embeddings. In this condition, the model received input which contained some sentences with three nested relative clauses and generalized to novel combinations of basic constructions with three nested relative clauses. This input to the model, however, reached beyond what we can reasonably expect to occur in child-directed speech—presumably such structures are completely absent in a child's ambient language. Moreover, this kind of structural generalization does not demonstrate the learnability of recursive productivity, because in training the model had experienced sentences with the same recursive depth as in testing.

6.7.5 Recursive generalization in the Dual-path model

To investigate whether the Dual-path model recursive capacities, I conducted an experiment in which the model's linguistic experience was limited to utterances with at most two nested relative clauses. The model was then tested on constructions with up to four nested relative clauses. An example of such a bizarrely complex sentence with four embeddings from the actual test set was:

(10) `the dog that was give -ing a stick that the woman that hit -ed the man that was jump -ing is present -ing to a girl to the mother is run -ing with a nurse .`

To enable recursive generalization in the model it was necessary to slightly change the meaning representations the model received compared to previous conditions in this chapter. For the current task, semantic features for all clauses had to be trained simultaneously, otherwise the model would with certainty not be able to produce novel embeddings correctly. In order to equally train all features in the event semantics, I randomized the mapping of features to clauses for all training sentences. This contrasted with previous experiments in which there was a fixed spatial relation between features and clause depth. In all other respects the message representation was identical to the TOPIC-FOCUS message used throughout. This change does not invalidate any of the earlier results, because it yields a semantics which is consistent with and more general than the previous semantics. On the contrary, independent tests not reported here indicated that many earlier generalization results could be improved upon with this semantics. It seems, however, that generalization in the model trades off against computational time, i.e., the message used here required more training episodes to learn the target language. For this reason alone I did not employ it in previous experiments.

The model was trained with this semantics on 10.000 message-sentence pairs randomly drawn from a language with at most two embeddings. The basic constructions from which these utterances were constructed were identical to the language described in Section 6.3. To assess the extent of recursive productivity in the model, it was then tested on 500 randomly generated, entirely novel sentences with three and four embeddings. Relative clauses in these test utterances could be attached to any admissible syntactic role and also relativize any syntactic. Thus, as the example item (10) suggests, relative clauses could be center-embedded or right-branching and could occur in any possible combination. Relative clauses in each sentence, however, were genuinely nested so that all novel test items had an embedding depth of either three or four. When the model was trained and tested in this way the results of Figure 6.9 were obtained (averaged over ten distinct training sets). The standard used to assess recursive generalization was *grammaticality.* This measure allowed minor errors in verb tense and articles but required that the grammatical structure of the target utterance was produced correctly. Figure 6.9 shows that all trained constructions—single-clause utterances and sentences with one or two embeddings—were learned to perfection at the end of training. Novel constructions with more relative clauses than experienced in learning were

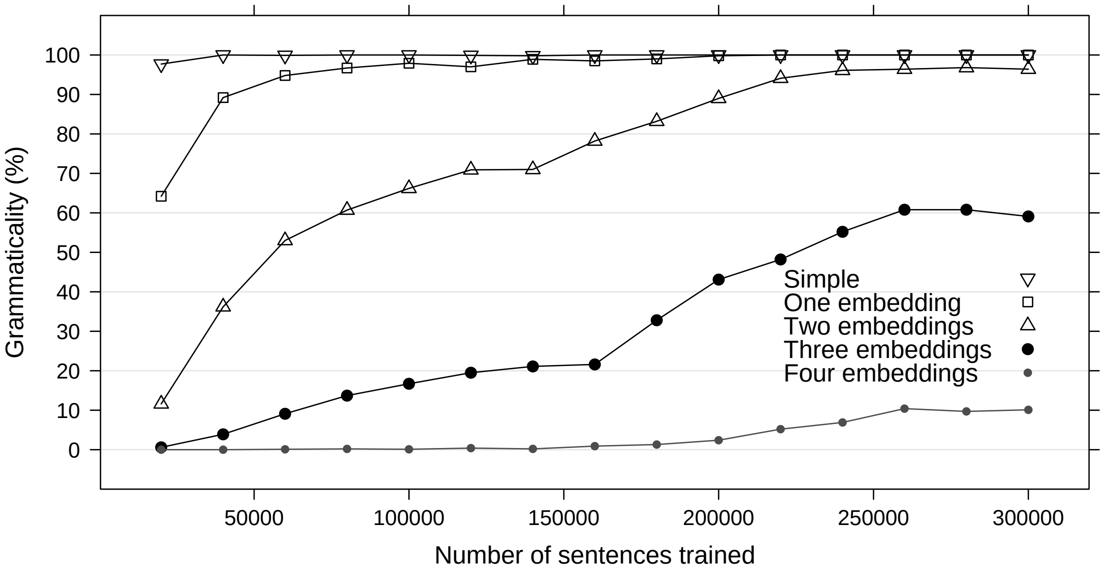


Figure 6.9: Recursive productivity in testing on novel complex utterances with three and four nested relative clauses.

produced with lower grammaticality. Utterances with three embeddings reached 60% grammaticality and utterances with four embeddings reached 10% grammaticality.

These results suggest that the Dual-path model is capable of significant recursive generalization beyond its immediate linguistic input. Although the model had not experienced any triple or quadruple embedded structure, it managed to grammatically produce the majority of test items with three relative clauses and still a non-zero amount of grammatical sentences with four relative clauses. Thus, the results also confirm that the model's performance rapidly degraded with the depth of embedding, which qualitatively matches human linguistic behavior. While the model's performance on novel triple embedded sentences was quite impressive, 10% grammaticality on novel sentences with four embeddings seemed rather low. Since there is little psycholinguistic data on the comprehension/production accuracy of such sentences, it is difficult to determine the closeness of fit with human processing. Recall, however, sentence (10) from above and how difficult it is to even judge grammatical.[23] Secondly, the fragment of the artificial grammar consisting of templates for utterances with four embeddings can generate roughly 4.8×10^{22} different sentences over the lexicon used. Statistically speaking, the model has therefore productively extended its knowledge of the target language by 4.8×10^{21} novel utterances through learning the syntax of simpler constructions. This achievement was measured in terms of grammaticality which compares actual utterances of the model with target utterances word-by-word. The average sentence length in the test corpus was 35.9 lexical items. Yet, if only one sentence position differed

[23] Miller and Isard (1964) report that their subjects were unable to learn sentences with three or four center-embeddings.

in word category from the target position, the utterance was discounted as ungrammatical. Often, the model produced test sentences with four embeddings which were almost grammatical and the performance improved as a function of training. To measure production accuracy in a more gradational way, I used a performance measure called *production error*, which was more sensitive to degrees of production success or failure than grammaticality. Production error measures the percentage of errors the model made out of all possible errors it could have committed. Subtracting this quantity from optimal performance yields the production accuracy score, which is a good measure of *performance*.[24] When the model's behavior was plotted in terms of production accuracy (Figure 6.10), the seemingly large difference (as measured by grammaticality) between trained constructions and novel test sentences diminished. Utterances with four relative clauses reached almost 90% production accuracy compared with the perfect score on trained constructions with at most two relative clauses. This is because production accuracy scaled the number of errors the model made by the length of the tested utterance.

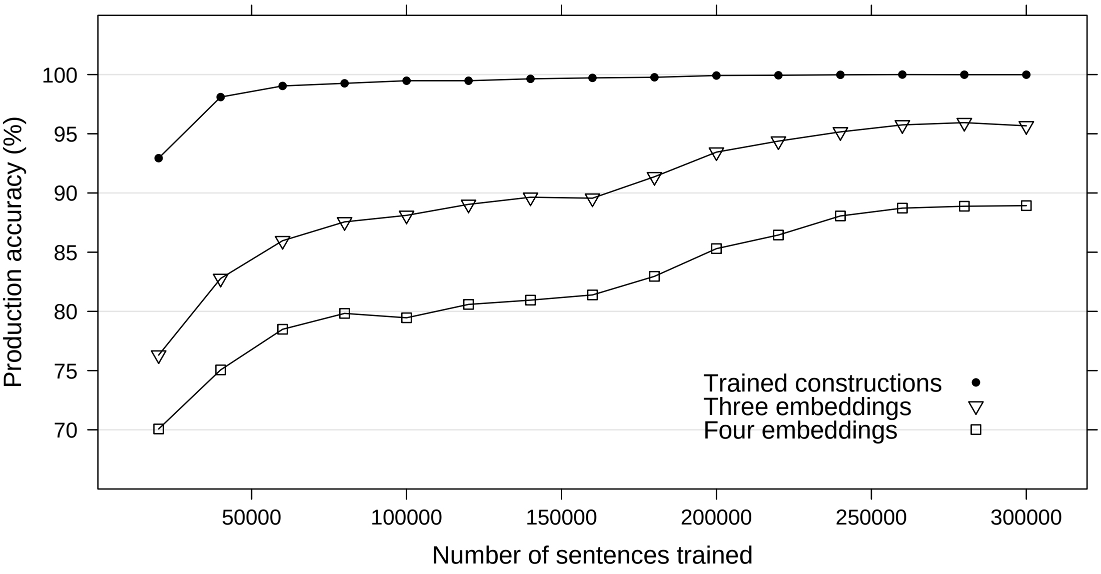


Figure 6.10: Production accuracy on novel third and fourth-degree embeddings, compared with mean over trained constructions (zero, one, or two relative clauses).

Based on these results I will now attempt to provide a preliminary answer to the questions (ii)–(iv) raised in 6.7.4 above. In order to acquire the syntax of trained constructions (with at most two embeddings), the model had to learn to appropriately se-

[24] Production error is based on a string metric called edit-distance, which compares target and output word-by-word. The edit-distance of two sequences of words is defined as the minimum number of primitive operations required to transform one sequence into the other, where primitive operations consist of an insertion, deletion, or substitution of one word, cf. Rohde (2002) for more details on this measure.

quence thematic roles for each clause type in the WHERE-layer according to the intended order. These construction types were signalled to the model by semantic features in the event semantics. Correct production of sentences with relative clauses required that the model learned a notion of relativization. This comprised the identification of the intended attachment site and the omission of the gapped element in the surface form. Learning both subtasks was enabled by the topic/focus features in the message-sentence pairs on which the model was trained. Furthermore, the model learned to associate different sets of semantic features in the message with different clauses in the sentence's hierarchical structure. This allowed the model to complete clauses before superordinate clauses were resumed (if applicable), or a new embedding started, without scrambling constituents from multiple clauses. Thus, the model developed representations which respected clause boundaries, clausal integrity, and the hierarchical organization of distinct complex constructions.

When tested on novel sentences with more levels of embedding than encountered in training, the message input as a whole was a novel semantic pattern to the model. Nonetheless, the model was familiar with message components at the clausal level and had the syntactic knowledge to combine several clauses into complex utterances by means of relativization. Unlike systematicity, which depended on the role-concept weights, recursive productivity depended on another part of the message, the event semantics. In training, the model learned to associate subparts of a sentence with the event semantics of the proposition that controlled it (Figure 6.11). The model learned from simple messages how to sequence participants in single-clause transfer events (`dog give toy to cat`). Other features of the event semantics controlled the position of relative clauses (`X that`) and the thematic role of the head noun in the relative clause (`that` *gap* VERB). When presented with a message for a novel construction, the model could use semantic regularities in the conceptual structure of the event semantics and combine these regularities to generate additional embeddings. From message-sentence pairs in training, the model learned which features of the event semantics controlled which aspects of the hierarchical organization of complex sentences. Since novel messages shared features in the event semantics with input messages, the model could generalize its learned subpart mappings and built novel structures from relevant message components. In this way, productivity was enabled by similarity-based meaning-to-form transduction.

There are, however, complications which can prevent this generalization. Sentences with three embeddings contain two clauses which are hierarchically 'sandwiched' between the main clause and the deepest embedding. This requires that the model determines the order of these two clauses by establishing the appropriate relations of co-reference between constituents in different clauses. Since conflicts of clause order did not occur in the training language and since none of these competing clauses was marked as more prominent in the message, the model needed to develop a policy to resolve such conflicts in the absence of training samples. To examine how the model achieved this, it helps to illustrate a condition in which the model *failed* to exhibit recursive productivity. In this experiment, I trained the model on an artificial language with at most one relative clause, trying to make it generalize to sentences with two and three relative

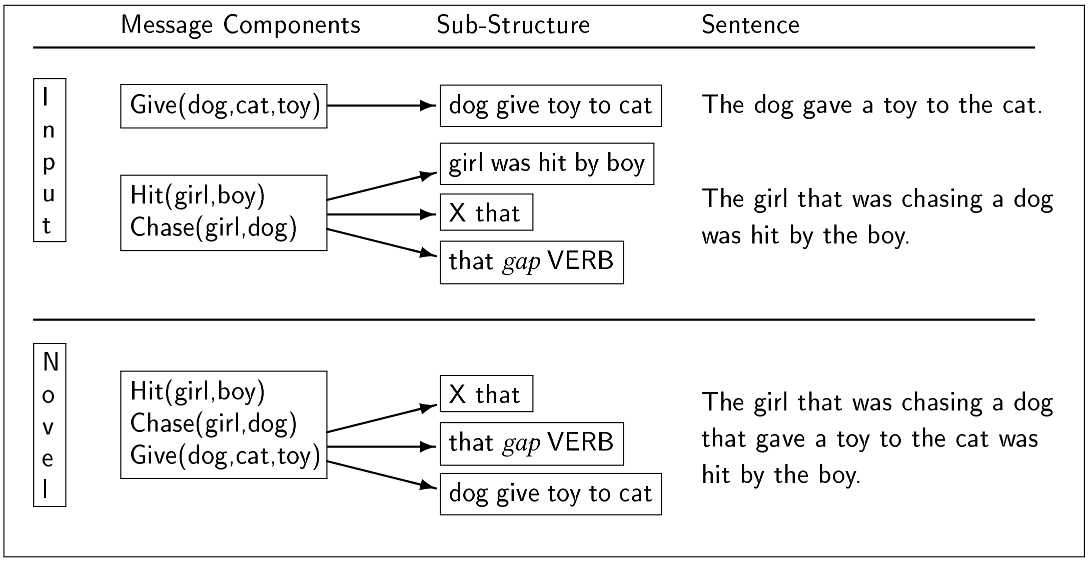


Figure 6.11: Different components of the message control different subsequences of words in the target structure.

clauses. The set up was identical to the experiment of Figure 6.9 in every other respect. In this condition, the model did not produce a significant amount of novel grammatical utterances with either double or triple embeddings. Hence, no recursive generalization to deeper embeddings occurred on the basis of exposure to sentences with one relative clause. This was surprising, given that generalization to two embeddings should be easier than to three or four. The model's output revealed that there was clause order confusion in virtually every tested sentence with two relative clauses. When trained on sentences with at most one relative clause, clause order is easily determined by the model because it can draw on the topic/focus features in the message. The topic feature always marks the main clause, the focus feature always marks the subordinate clause. Test sentences with deeper embeddings, however, had at least two such features each and the model had not learned how to negotiate the resulting conflict. When trained on sentences with two relative clauses, on the other hand, the model developed a strategy to 'chain' clauses in the correct order and this strategy transfered to test items with deeper embeddings. An isolated topic feature signaled the main clause to the model. Once the head was produced and fed back to the model, activation spread to the corresponding concept in the CWHAT-layer. This activated the head noun's thematic role in the CWHERE-layer, but also the thematic role of the gapped element which was linked to the same concept. Since the CWHERE-layer projected into the HIDDEN-layer and the gapped role was clause-specific, the model could use this cue to determine which clause was to be sequenced next. This process repeated for the following embedding, and so forth. In other words, deeper embeddings in the input forced the model to attend to subtle cues in the message-lexical pathway in order to sequence clauses of the intended structure. Since the sketched strategy is generic, it applied in recursive generalization beyond the model's input as well. When these more complex structures were absent from the learning environment, there was no need for the model to develop such a strategy which is why it failed to generalize recursively.

Generalization to deeper embeddedings can be described as a special instance of sequence learning—from exposure to the syntax of zero, one, and two embeddings to the syntax of three or more embeddings. This task involves recognizing structural regularities in the observed sentences of limited depth and making syntactic predictions for novel sentences with deeper embeddings. To answer question (ii), the Dual-path model could accomplish this task because of its (a) sensitivity to semantic similarities between trained and novel messages at the clausal level, because (b) it learned an abstract notion of relativization which transferred to novel constructions, and because (c) it developed a generic policy for ordering clauses. The combination of such knowledge, acquired from the linguistic input, might therefore be sufficient to explain recursive productivity in humans. When the sequence of syntactic structures that the model was exposed to was limited to one embedding, relative clause constructions in the input were perceived as idiosyncratic. The learning sequence, or 'inductive basis', for syntactic generalization was too small. Without linguistic evidence that relativization can be iterated at least once within a relative clause itself, the Dual-path model did not generalize recursively. It would be an empirical prediction of the model that based on exposure to sentences with at most one embedding, human learners cannot achieve recursive productivity either.[25]

Recursive productivity in the Dual-path model degraded with the depth of embedding (Figure 6.9, page 191). There might be several factors responsible for this behavior. One factor might be sentence length. Longer sentences create longer dependencies, which might tax the working memory of the processor. In Section 6.7, I argued that working memory limitations did not adequately explain the model's differential performance on right-branching and center-embedded sentences. The number of errors characteristic for working memory limitations did not increase with the distance between dependent constituents. At the same time, such errors were not completely absent. Hence, working memory might still play a minor role in the model's behavior. Another factor might be that sentences with more clauses contain more *semantic* dependencies which have to be encoded in the model's message. Figure 6.12 on page 196 shows some of the complex semantic relations in test sentence (10) with four relative clauses (page 190). Some of these relations are signalled to the model by semantic features and their relative level of activation, some relations (such as co-reference) have to be inferred by the model during processing. In general, more clauses entail more semantic relations in a sentence which entails more message features being involved in representing sentence meaning. All message features are concurrently active from the beginning of production, hence the more clauses a target utterance has, the more features are active. This

[25] Artificial grammar learning would perhaps be a suitable paradigm to test this claim. Intuitively, it is obvious that sequence learning from $s_1, s_2, \ldots, s_j$ to predicting s_{j+1} cannot succeed if j=2 because no tentative prediction from s_1 and s_2 can be tested *within* the learning sequence.

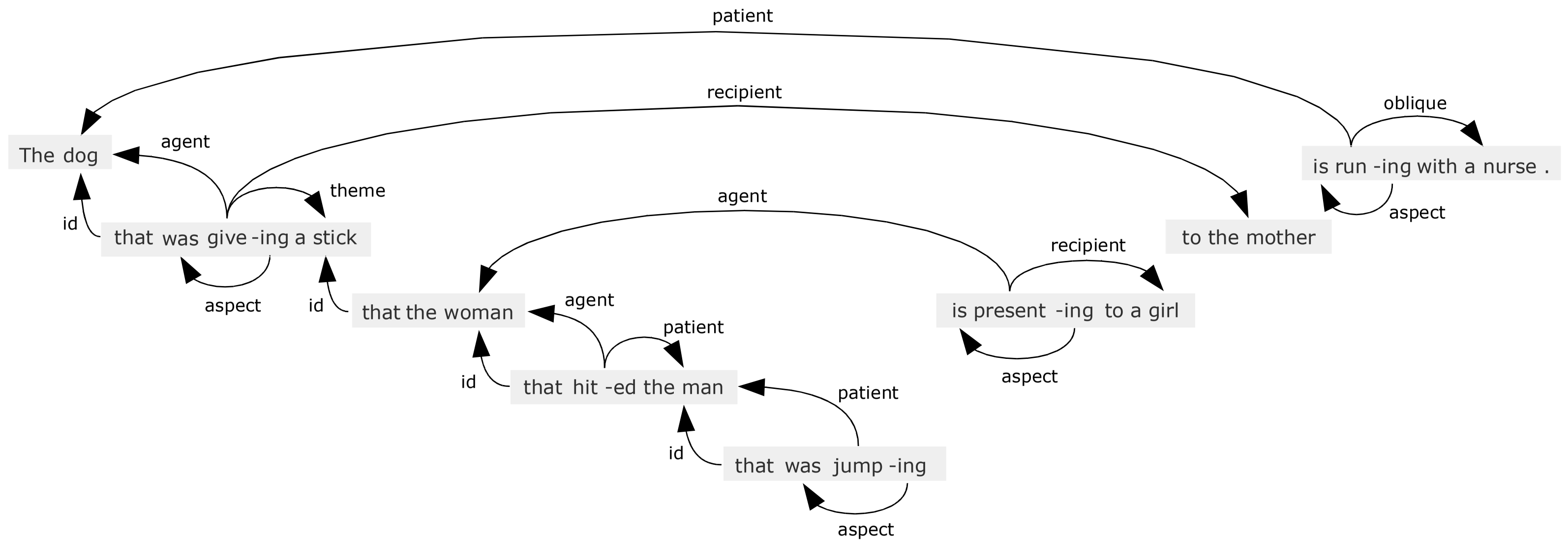


Figure 6.12: Semantic dependencies in test sentence (10) signalled by or inferable from the message input to the model. The co-referentiality of head nouns and gapped elements is omitted for clarity, 'id' abbreviates 'identifies'.

information has to be processed by the HIDDEN-layer and utilized sequentially at the right position. But the HIDDEN-layer is of fixed size and consequently there are limits to the amount of information it can process effectively. Neural networks such as the Dual-path model are sensitive to activation differences between units, but only to a certain extent. If HIDDEN-layer activation states become too similar for distinct inputs, they will not lead to distinct output anymore. With deeper embeddings and more semantic information exciting the HIDDEN-layer units, differences in the message might go undetected, leading to more errors in the model's productions.

A third factor in the performance degradation lies in the way the model represents syntactic knowledge over one set of connection weights. Although the HIDDEN-layer develops specialized units or assemblies of units to represent specific aspects of syntax, these are not discrete components which are functionally independent, because every feature in the EVENT SEMANTICS-layer projects into every HIDDEN-layer unit. Hence, additional features in test items with more embeddings can disturb specialized assemblies and cause less distinct or even disruptive patterns of activation at the WHERE and COMPRESS-layer. To illustrate this point, consider the simple feed-forward network of Figure 6.13 with a firing threshold of 1 for each unit. Each subnetwork, $\mathrm{SN}_1 = \{i_1, i_2, h_1, h_2, o_1\}$ and $\mathrm{SN}_2 = \{i_2, i_3, h_1, h_2, o_1\}$, is implementing the logical XOR-function.[26] However, if input unit i_3 (i_1) is active this will destroy the XOR behavior of SN_1 (SN_2) in case $i_1 = i_2 = 1$ ($i_2 = i_3 = 1$). The input units $\{i_1, i_2, i_3\}$ of this network can be conceived of as features in the event semantics of the Dual-path model, the hidden units $\{h_1, h_2\}$ as HIDDEN-layer units, and the output unit o_1 as a COMPRESS-layer unit. Suppose the model has been trained on sentences with two relative clauses and has developed a policy to deal with a pair of input features resembling the XOR-function. In testing the model on utterances with more embeddings a third feature is activated in the message which projects into one XOR-subsystem. This additional feature might then interfere with the processing of semantic relations in other clauses of the test utterance similar to the disruption of the XOR-behavior. As a consequence, activation is erroneously propagated to the COMPRESS-layer, which represents abstract syntactic frames. This causes the model to report wrong syntactic choices to the WORD-layer (e.g., by activating an incorrect word category) and produce inaccurate sentences. The more clauses (and thus semantic features) a novel test utterance has,

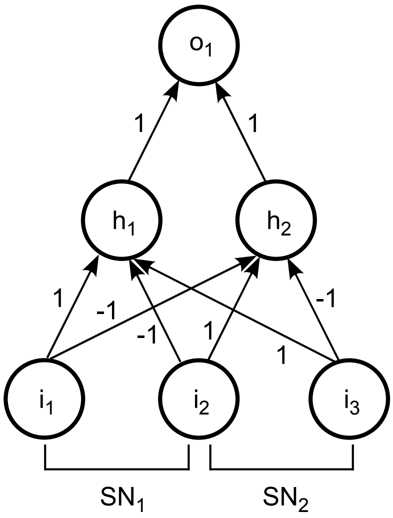


Figure 6.13: A feed-forward network with two subnetworks each implementing the XOR-function.

[26] $\mathrm{XOR}(x, y) = 1$ iff $x \neq y$ with $x, y \in \{0,1\}$.

the more such interference might occur and the less accurate the model's production becomes.

If this explanation is correct, we should observe an increasing amount of intrusions of spurious activation at the COMPRESS-layer with increased depth of embedding. To test this, I recorded the activation states of the COMPRESS-layer in the way described in Section 6.7 while the model processed genuinely right-branching test utterances with up to four relative clauses. The overall picture resulting from this procedure was that the number of completely inactive units decreased with the depth of embedding suggesting that patterns of activation became less and less distinct at the COMPRESS-layer. As an example, I plotted the activation states of unit C25 which was identified earlier as encoding syntactic differences in the hierarchical organization of sentences (Figure 6.14). Four sentences were picked which shared the same initial segment of word categories (one sentence for each of four levels of embedding). While trained structures with one and two embeddings showed a clear-cut pattern, with all units either active or silent, there are first intrusions of activation in the triple embedded structure and even more perturbing activation in the test utterance with four relative clauses. This blurring of activation differences caused the model increasing difficulty to map its message input onto the syntactically correct sentence form. It is important to point out that less distinct patterns did not result from a failure to learn; the model was not exposed to sentences with four embeddings and it learned two embeddings to perfection. Nor do these patterns stem from erroneous word feedback because the model produced all plotted sentences correctly. Hence, they must result from additional semantic features in the message of novel utterances, interfering with learned procedures for sequencing clauses in simpler structures. This interference leads to a degeneration in performance with depth of embedding since deeper embeddings entail more semantic information in the message.

Sequence of word categories	COMPRESS-layer unit C25 by depth of embedding			
	1	2	3	4
NOUN				
VERB			░	
NOUN				
THAT	■	■	■	■
VERB				
NOUN				▒
THAT		■	■	■
VERB				░
NOUN			░	▒
THAT			■	■
VERB			░	▓
NOUN				▒
NOUN				▒
THAT				■
VERB				▓
NOUN				▒
NOUN				░

Figure 6.14: Activation of COMPRESS unit C25 for right-branching sentences with up to four relative clauses.

To summarize, the Dual-path model's recursive productivity is degrading with increased sentence complexity and this property can be explained by multiple factors including sentence length, the number of semantic dependencies in the sentence, and interferences from novel combinations of message features. No stipulated constraints on the language processor are necessary to account for this behavior. The proposed explanation of degrading performance is broadly consistent with recent interference-based theories of complex sentence processing in humans (Gordon et al., 2001; Lewis and Vasishth, 2005).

In response to question (iv)—can *gracefully terminating recursive productivity*, be achieved without learning explicit rules of recursive composition?—the answer should be fairly obvious: yes, this has just been demonstrated. It might be objected that such rules are merely 'hidden' but not absent in connectionist networks (see the debate between Marcus (1999a,b) and Seidenberg and Elman (1999a,b)). There are two aspects to this kind of objection, the training environment of the Dual-path model was generated using explicit rules and the model was trained with feedback from a 'teacher' on items from this environment. Weights in the model are adjusted according to behavioral mismatches with the recursively generated training patterns. Therefore, it might be argued, recursive rules become a part of the model itself during learning; the model does not eliminate rules but rather implement such rules when viewed as integrated system consisting of the training items, the network and its control structure. Such objections, however, are misguided in several ways. First, feedback links the structural properties of the training set with model behavior, but generating feedback does not require the application of recursive rules. The model is taught in a word-to-word prediction paradigm in which every word output is compared with the desired target. Thus, feedback results from a local comparison of word positions, and does not inform the model about the grammaticality of a complete output sequence. And secondly, through such feedback, the model is not instructed to generate sentences in the same way as the rule-based, artificial language generator. The training signal carries no information whatsoever about these rules of generation, although it does carry information about the extension of these rules. Consequently, the model is not taught the recursive rules involved in generating the training environment, it is taught to behave *as if* it implemented such rules. A third point of confusion concerns the notion of implementation itself.[27] A connectionist system such as the Dual-path model might be describable as implementing rules of recursive syntax. But there is still an ontological difference between systems which at some level of description follow linguistic rules in processing and systems which are rule-based. The difference is that in the latter but not the former systems rules are *causally efficacious* components. In this sense at least, the Dual-path model approach to recursive productivity eliminates syntactic rules in complex sentence processing.

[27] Elsewhere I argued that the concept of implementation is not well-defined and in need of explication in order to avoid fruitless debates in physical and biological computation (Fitz, 2007).

6.8 Conclusion

Virtually everyone involved in the study of language can agree that grammars are mappings between meaning and phonological form. As Pullum and Scholz (2008) remark, however, rarely does anyone seem to take this commonplace seriously in how grammar is modelled. Certainly not the mainstream of theoretical linguistics, where grammar is a system to generate sentences, although the view of grammar as a conventionalized meaning-to-form mapping is a central tenet of Cognitive and Construction Grammar (Taylor, 2002; Goldberg, 2006). In these theories, grammar is not an autonomous device which interfaces with semantics and phonology but a stored inventory of direct associations between meaning and form. A system for language processing has to learn such pairings of meaning and form by *transducing* phonological representations into meaning representations in comprehension and vice versa in production. The Dual-path model implements transduction in a straightforward way in that it maps message input onto appropriate sentence forms. By learning from message-sentence pairs the model develops internal representations which constrain this mapping. These constraints can be viewed as the grammar of the target language, having evolved from successfully learning to transduce between types of representations. Thus, the grammar acquired by the Dual-path model is not a means for building syntactic representations which are then imbued with semantic or phonological content, grammar in the model associates such content in unmediated transduction.

In the experiments described in this chapter, I found modelling grammar as a fall-out product of meaning-to-form transduction to be a quite powerful approach to linguistic generalization. Unlike other neural network models which have been tested on their ability to learn complex sentence structure (e.g., the SRNs of Elman (1991) and Christiansen and Chater (1999b)), the Dual-path model is not inducing grammar from sequences of words but develops a grammar in learning to map semantic content onto sequences of words. In this process the model draws on statistical regularities in the sentence input (e.g., frequencies of lexical co-occurrence) but also on semantic information in the message, such as role-to-concept bindings and the event structure of intended utterances. As the model learns to interpret and utilize this information, transduction becomes increasingly accurate to the point where the model's grammar reflects all meaning-form pairings of the trained language. Generalization occurs because of semantic similarities between constructions encoded in the message input. Novel utterances are produced correctly to the extent that they are relevantly similar in meaning to experienced utterances. I argued throughout this chapter that similarity-based transduction can explain structural generalization, strong systematicity, and recursive productivity in the Dual-path model.[28] These generalization capacities and their characteristic behavioral profiles testify to the model's potential as a suitable model for human language processing.

Generalization is enabled but not guaranteed by or reducible to semantic similarity

[28] For the remainder of this thesis I will therefore refer to the Dual-path model, modified for the processing of multi-clause utterances, as the *recursive Dual-path model*.

between experienced and novel utterances. Whether the model generalizes also depends on the strength of learned meaning-form associations, i.e., on distributional properties of the learning environment. This factor was clearly identifiable in Section 6.7 on the differential processing of right-branching and center-embedded constructions. Furthermore, generalization of course depends on the model's transduction grammar, i.e., on the type of constructions experienced in learning. In Section 6.7 it was argued that the model failed to be recursively productive in an impoverished input condition. Semantic representations were identical across conditions, but the model was 'blind' to similarities in the message.

Representations were hand-coded and feature-based similarities imposed on the model by the experimenter may not yield optimal generalization behavior in different input conditions. It would therefore be a worthwhile future project to find semantic representations which maximize generalization capacities for changing learning environments, e.g., through evolutionary programming. Based on these findings, it might be possible to define a metric of semantic similarity which predicts differential generalization and to test the sensitivity of human learners to this metric in sentence processing.

Chapter 7

Learning polar interrogatives

In this chapter I exploit the generalization properties of the recursive Dual-path model to provide evidence for the data-driven learnability of complex polar interrogatives. I will argue that the model favors *structure-dependent* over *structure-independent* auxiliary fronting and identify learning conditions in which it can produce correct complex polar interrogatives in the absence of positive exemplars of these structures in the input. The model's behavior is matched against child language data and compared with other approaches to complex question learning. Since the model does not implement a traditional kind of language-specific universal grammar, these results are relevant to the *poverty of the stimulus* debate.

7.1 The general controversy

The acquisition of natural language is a complex process in which children learn to comprehend and produce utterances of their native speech community from the ambient language in their social environment. One of the most persistent controversies in cognitive science concerns the question whether this task can reliably be accomplished based on the linguistic input the child receives during the ’critical period’. Given some property $\mathcal{P}$ of a language $\mathcal{L}$, can $\mathcal{P}$ be known through sensory experience alone, or do we need to posit language-specific sources of information other than the ’primary linguistic data’ to account for the fact that children eventually come to know $\mathcal{P}$? Many syntactic properties are abundantly warranted in the linguistic input for children to acquire knowledge of $\mathcal{P}$ from experience. Such $\mathcal{P}$ include word-order, branching-direction, case marking, and the morphosyntax of tense and aspect. Although word order, for instance, differs across languages, most languages have a preferred or dominant word order which can be learned from the structures in the linguistic input. In other cases, however, the experiential basis to infer $\mathcal{P}$ seems considerably weaker. Prime examples of such $\mathcal{P}$ (in English) are subjacency constraints on forming complex *wh*-questions (Chomsky, 1986), ‘want-to’ contraction (Crain and Thornton, 1998; Crain and Pietroski, 2001), and the assignment of nominal antecedents to anaphoric *one* (Baker, 1978; Lidz, Waxman, and Freedman, 2003).[1] These properties differ from, say, word order in that they involve a generalization which has exceptions. The verbal elements *want* and *to* can be contracted into *wanna* in some constructions such as:

(1) a. *Who does he want to cook dinner for?*
b. *Who does he wanna cook dinner for?*

In other constructions, however, contraction is not permissible:

(2) a. *Who does he want to cook dinner?*
b. * *Who does he wanna cook dinner?*

In acquisition, children have to form a generalization $\mathcal{P}$ = "*want* and *to* can be contracted unless X" rather than $\mathcal{P}'$ = "*want* and *to* can be contracted". While $\mathcal{P}'$ could be learned from relevant examples, it has been argued that correct contraction $\mathcal{P}$ is not learnable from experience because of the limited availability of input which restricts $\mathcal{P}'$ appropriately (Crain and Pietroski, 2001). Similarly, subjacency and anaphoric *one* are presented as problematic in the literature, because children could not retract from overgeneralization. The unconstrained generalization $\mathcal{P}'$ is warranted by the linguistic input, but the more complex property $\mathcal{P}$ is underdetermined by experience. The acquisition problem is further exacerbated in that child-directed speech is non-uniform. Constraining evidence might be available to some children but not others. And yet all normally developing children rapidly converge towards knowledge of $\mathcal{P}$ regardless.

The contraction example illustrates the view that there are syntactic properties $\mathcal{P}$ which are neither sufficiently nor reliably supported by a learner's environment. In other words, there is a gap between the information provided by sensory experience of $\mathcal{L}$ and what children end up knowing about $\mathcal{L}$. This relational deficit between learning and experience is often referred to as the *poverty of stimulus.*[2]

Properties $\mathcal{P}$ of $\mathcal{L}$ for which such a gap exists, harbor an explanatory problem given that children eventually acquire adult-like knowledge of $\mathcal{P}$. This *explanandum* has been labelled 'hyperlearning' (Pullum, 1996): how do children invariably settle onto knowledge of $\mathcal{P}$ in the absence of sufficient evidence (either positive or negative) to isolate $\mathcal{P}$ from its competitors? Instances of hyperlearning seem to necessitate a popular doctrine—linguistic nativism—which stipulates the innate guidance of learning through language-specific control mechanisms (universal grammar). If knowledge of some $\mathcal{P}$ does not derive from sensory experience of sentences in $\mathcal{L}$, it is best explained *from within.* This line of reasoning is exemplified in Lidz et al. (2003) who contend that knowledge of (some) syntactic properties "must derive from linguistic structure inherent in the learners themselves because [...] the input to which infants are exposed does

[1] For more examples confer the list in MacWhinney (2004).

[2] *Poverty of stimulus* is an umbrella term for a variety of different claims, including the degeneracy of the input, the unavailability of corrective evidence, the inductive quandary of finite input versus productive infinity, and the formal non-learnability results of Gold (1967).

not unambiguously support the linguistic representations that they create" (p. B72). Nativist arguments are strengthened when hyperlearning occurs in the absence of mistakes during syntactic development ('error-free learning', MacWhinney, 2004).

There are several strategies to avoid the nativist "solution". One can contest that children unerringly arrive at knowledge of $\mathcal{P}$ by providing developmental data to the contrary. This approach undermines the explanatory value of universal grammar, but it does not answer how knowledge of $\mathcal{P}$ can be established from deficient input. Secondly, it could be demonstrated that the primary linguistic data is sufficiently rich to obtain knowledge of $\mathcal{P}$ from experience by showing that corpora of child-directed speech contain more relevant samples which constitute direct evidence for $\mathcal{P}$ than claimed. In this case, there is no *poverty of the stimulus* and hence no need to explain hyperlearning. Another way to approach the learning problem is to argue that the empirical basis of a child is broader than supposed in that it comprises linguistic structures which are not paradigm cases but nonetheless aid the acquisition of $\mathcal{P}$ in relevant ways. Learning in many domains is *path-dependent*, i.e., one cannot get everywhere from just any state of knowledge. There may be a path to $\mathcal{P}$ which opens up once other pieces of knowledge about $\mathcal{L}$ have been established through experience. If the child's learning mechanism draws on wider resources in the right order, the data-driven learnability of $\mathcal{P}$ could become plausible even though pertinent samples of a syntactic structure may be highly infrequent in the linguistic environment. And finally, one might accept that in a strict sense the primary linguistic data is impoverished (i.e., some target structure exemplifying $\mathcal{P}$ is largely absent) and that innate learning biases are required to explain the fact that children reach adult-like knowledge of $\mathcal{P}$ with high probability. Yet, it need not be assumed initially that these biases are language-specific. It could be argued that children's input filtered through the architecture of a processor which incorporates domain-general learning biases of some sort is sufficient to obtain knowledge of $\mathcal{P}$. Ideally, such an argument would be based on a computational model of syntactic development which can demonstrably and reliably acquire knowledge of $\mathcal{P}$ from noisy, realistic distributions.

I will now look more closely at a specific linguistic structure that has often been cited as a paradigmatic example of the poverty of stimulus—yes/no-questions with relative clauses. The controversy over the learnability of these questions forms an important strand in the *nature versus nurture* debate since several decades. The strategies and approaches described above have all been taken in this debate and will be illustrated in more detail at work.

7.1.1 The learning problem

The issue of interest here is whether normally developing children can arrive at syntactic knowledge of $\mathcal{P}$ in $\mathcal{L}$ which cannot conceivably have been extracted from the speech they were exposed to. The strongest and most frequently adduced case of such knowledge concerns auxiliary fronting in English polar interrogatives. Declarative sentences can be transformed into yes/no-questions by inverting the positions of the sentence initial subject NP and the verb auxiliary. Suppose in acquisition a child encounters the declarative sentence

(3) *The dog is barking.*

together with examples of yes/no-questions

(4) *Is the dog barking?*

From a purely logical point of view there are many hypotheses about the underlying rule of grammar a child may come to entertain which are compatible with this linguistic experience, for instance, *'move the third word in front'* or *'place the left-most auxiliary in front'*. For single-clause sentences such as (3), which contain only one auxiliary, the latter rule of question formation is descriptively adequate, but only incidentally. It is invalidated by declaratives which contain multiple clauses with auxiliaries, such as

(5) *The dog that is chasing the cat is barking.*

Placing the left-most auxiliary in front yields the ungrammatical question

(6) **Is the dog that chasing the cat is barking?*

instead of the grammatical form

(7) *Is the dog that is chasing the cat barking?*

Multi-clause sentences such as (5) show that it is the auxiliary of the *head* (main clause) which is moved to the initial position and not, e.g., the sequentially first auxiliary. Linguists call such transformations *structure-dependent* to emphasize that they are governed purely syntactically. That is to say, the choice of which auxiliary is placed in the initial position does not depend on the linear order of constituents but only on the hierarchical organization of the sentence. The structure-dependence of auxiliary fronting in English was first pointed out by Chomsky (1965) and he connected polar interrogative learning with the insufficiency of the primary linguistic data. According to Chomsky, complex questions such as (7) are virtually absent from child-directed speech. This claim is expressed in Chomsky's notorious statements that "a person might go through much or all of his life without ever having been exposed" to complex yes/no-questions and that "you can go over a vast amount of data of experience without ever finding such a case" (see Piattelli-Palmarini, 1980). Moreover, Chomsky argued that simple questions like (4), which are quite frequent in child-directed speech, support a *structure-independent* rule because in both cases the auxiliary which is closest to the subject NP is placed in front. Hence, children should form an erroneous generalization, and since they rarely, if ever, hear sentences of the appropriate sort—center-embedded interrogatives of type (7)—they will not be able to retract from it based on the sentences they hear. Because linguistic experience is impoverished in this way but children learn structure-dependent auxiliary fronting nonetheless, Chomsky (1975) suggested "the only reasonable conclusion is that UG contains the principle that all such rules must be structure-dependent". The view that structure-dependent knowledge must be innate has since been endorsed by many linguists and psycholinguists (e.g., Crain and Nakayama, 1987; Crain and Pietroski, 2001; Legate and Yang, 2002).

7.1.2 Three empirical hypotheses

For the sake of argument, let's assume that Chomsky's claim regarding the poverty of the stimulus is true of child-directed speech with respect to polar interrogatives; I will summarize the empirical evidence for this claim below. Then, structure-independent rules of auxiliary fronting are consistent with the learner's input and readily available from exposure to single-clause questions. Hence, such rules would be inductively supported and because they arguably are simpler than the correct rule, we would expect children's early productions to reveal errors deriving from the wrong choice of auxiliary fronting rules. Surprisingly, however, children do not seem to select any of the incorrect alternatives, not even temporarily. Crain and Nakayama (1987) conducted an experiment in which complex polar interrogatives were elicited from children aged [3;2]–[5;11]. Subjects were requested to ask questions to a doll, e.g., *"ask Jabba, if the boy who is watching Mickey Mouse is happy?"*, and these elicited sentences were scrutinized for errors. As their main result, Crain and Nakayama (1987) found that no child produced complex yes/no-questions in which the auxiliary of the relative clause was fronted. All of their subjects seemed to entertain a structure-dependent hypothesis and none had overgeneralized to an incorrect structure-independent rule.[3] Consequently, according to this study, children do not commit structure-independent errors in syntactic development which then require correction at a later stage; auxiliary fronting appears to be a case of error-free learning.[4] It must be pointed out, however, that this study is neutral with respect to the question whether children develop knowledge of syntactic structure *without* being exposed to the relevant kind of evidence that could warrant such knowledge. In other words, while auxiliary fronting in polar interrogatives might be an example of error-free learning, it has not been established that it is a case of hyperlearning.

One can take error-free learning as evidence for innate language-specific knowledge (as Crain and Nakayama do), but one might just as well conclude that the very learning problem for auxiliary fronting in its standard rendition of Section 7.1.1 is phrased in a misleading way. The problem was formulated as a choice between two principles and

[3] This is not to say that these children did not make any errors in this task. I will come back to this issue later when comparing production errors of the recursive Dual-path model with children's error profiles in the study of Crain and Nakayama (1987).

[4] The methodology of this study can be criticized, because the correct relative clause (without auxiliary displacement) was provided by the experimenter in the elicitation instruction and could be imitated by the subjects to yield correct questions. Furthermore, Thomas (2002) remarked that these results "only showed that children don't produce structure-independent questions; it didn't prove that children reject [them] as ungrammatical" (p. 67). For an extensive review and criticism of the Crain and Nakayama (1987) study, confer Ambridge, Rowland, and Pine (2008).

it was presented as a puzzle of how children could avoid the incorrect and arrive at the correct generalization. But if children do not make mistakes which could be attributed to an incorrect generalization, then they apparently never entertain overgeneralizations which derive from simple polar questions in the input. In fact, if children never witness a complex polar question, as the nativist argument assumes, why would they project *any* syntactic hypothesis about such structures in the first place? Similarly, if children never witness a construction in which an auxiliary is extracted from a relative clause, why would they ever entertain such a rule for polar questions? The formulation of the learning problem presupposes that inductive generalization from simple to complex polar questions would be the natural and obvious learning strategy for a child. But the data of Crain and Nakayama suggests that this assumption is unwarranted, otherwise we would observe characteristic overgeneralization errors. If syntax learning does not proceed by inductive generalization, there is no inductive underdetermination in choosing between competing hypotheses and the innateness of structure-dependence would seem to be a solution to a learning problem which might not exist in the sketched form. The question how auxiliary fronting is acquired, however, does not go away just because the logic of the 'classical' formulation of the problem might be flawed.

Nativist accounts of auxiliary fronting might rest on a misguided conceptualization of the learning problem as an inductive choice. A number of further assumptions seem to be implicit in the claim that children are unable to infer the correct fronting rule from linguistic experience:

(i) polar questions are formed from declaratives by transformational rules over strings of words, and these rules are the child's learning target.

(ii) children require a critical amount of positive samples of complex polar interrogatives of type (7) to learn yes/no-question formation, and this amount is not available.[5]

(iii) only complex polar interrogatives of type (7) are relevant to learning yes/no-question formation.

Furthermore, the nativist account assumes that

(iv) innate priming *best explains* error-free acquisition data such as Crain and Nakayama's.

In computational terms, learning problems involve a target domain, a learning mechanism, and an information source. Trivially, the success of a learning system will depend on the specification of these components (and the criterion of success). In the above formulation of the task, the learner must decide between competing hypotheses in the form of monolithic syntactic rules which is claimed to be impossible without innate priming. This inductivist approach contrasts with a constructivist approach in which

[5] To put it differently, even if there were some complex polar interrogatives in the input, children would not be sensitive to them.

the syntax of polar questions is assembled piecemeal from experience. If we reformulate the learning target in terms of simpler building blocks, it may become tractable: (a) learn the syntax of single-clause yes/no-questions (subject/auxiliary inversion), (b) understand the function of relative clauses, (c) conjoin this knowledge to produce correct multi-clause yes/no-questions.

In poverty of the stimulus arguments it seems to be presupposed that the acquisition of polar questions precedes the acquisition of hierarchical clause structure. A study on syntactic development by Diessel and Tomasello (2005) shows, however, that children of the average age in the Crain and Nakayama (1987) study master various relative clause constructions quite well as measured in sentence production (see also Kidd et al., 2007). In fact, it is a methodological necessity in the Crain and Nakayama experiments that children understand the request to produce questions which contain a relative clause. It can therefore be assumed as plausible that children in these experiments had an understanding of the function of relative clause constructions as modifiers, to fix a topic and/or referent, or to provide additional information. As a consequence, they have some knowledge of clause boundaries, clausal dependencies and hierarchical sentence structure. Furthermore, it can be assumed that children know about the pragmatic function of simple yes/no-questions based on the primary linguistic data. On Chomsky's view the correct transformational rule is inaccessible to children, but if such knowledge is brought to the table, it would rather be in need of explanation why relative clauses *should* interfere with auxiliary fronting. Supposing that clausal units and dependencies are recognizable by the child and that different units serve different communicative functions, the structure-dependent principle may be learnable precisely because it is structure-dependent.

In similar vein, Van Valin (1998) proposed a pragmatic motivation why the identification of additional information may prevent a child from fronting the wrong auxiliary:

> Questions are requests for information and the focus of a question signals the information desired by the speaker. It makes no sense, then for the speaker to place a focus of the question in a part of a sentence which is presupposed, i.e., which contains information which the speaker knows [...]. The content of adverbial clauses and restrictive relative clauses is normally presupposed, and consequently constructing questions with the focus in one of these structures generates a pragmatic contradiction. (p. 232)

Based on such pragmatic considerations Van Valin formulates a general restriction on the formation of simple yes/no-questions and extends this principle to the more complicated *wh*-questions. In this manner, the learnability problem for subjacency constraints on *wh*-questions, another prime suspect for the poverty of stimulus, is reduced to acquired knowledge of forming yes/no-questions with embeddings. According to Van Valin, this learning task can in turn be grounded in experience of simple yes/no-questions plus prior semantic information and pragmatic constraints.

In a nutshell, then, instead of learning a movement rule by direct observation, children may be capable of assembling the syntax of complex yes/no-questions from more basic principles of question formation and relativization:

> Hypothesis-1: *Complex polar interrogatives can be learned from simple polar interrogatives and relative-clause constructions in the absence of positive exemplars in the input.*

I will test this hypothesis in my computational model of syntactic development in Section 7.2. Hypothesis-1 is rooted in doubt about whether the learning task for polar interrogatives—assumption (i)—is adequately described in standard formulations. If Hypothesis-1 can be validated, this would also cast doubt on assumptions (ii) and (iii), viz that some complex polar interrogatives are required in the input and that the target syntax could not be arrived at in other ways, e.g., by piecemeal, bottom-up construction from simpler building blocks.

Assumption (ii) suggests that *if there was* sufficient and unambiguous evidence, children *could* learn the proper syntactic rule for auxiliary fronting. Purely data-driven learning could succeed if only there was enough of the right kind of examples present in child directed speech. In this case, there would be no inductive or *logical problem* involved in learning structure-dependent principles. Whether positive forms are sufficiently frequent to rule out competing syntactic principles is a downright empirical question. But what is the right kind of evidence and how much of it would be sufficient? Pertinent information is conspicuously absent from the nativist literature, apart from flat denials that there is any such evidence. Pullum (1996) attempts to debunk these claims as unwarranted. He first argues that there is more evidence for the correct generalization than supposed (see also Sampson, 1989). According to Pullum, it is not merely polar questions like (7), but also other types of questions such as

(8) *If you're done with eating, could I have your french fries?*

and even *wh*-questions such as

(9) *Why couldn't anyone who was at home close the window?*

which constitute evidence for the correct auxiliary fronting rule. In forming questions (8) and (9) there is a similar structure-dependent auxiliary movement involved as in (7). Placing the first auxiliary of the corresponding declaratives in sentence-initial position would result in ungrammatical questions:

(10) **Was anyone who at home could close the window?*

Thus, Pullum suggests that learning complex polar interrogatives is supported by a larger variety of positive examples in the primary linguistic data; any observed main clause auxiliary displacement in utterances with subordinate or complement clauses might be relevant. Although target structures such as (7) may be highly infrequent, it is not too far-fetched to expect such mundane questions as (8) and (9) to occur in child-directed speech. To determine the frequency of such expressions, Pullum examined the

Wall Street Journal corpus in the Penn TreeBank and suggests that it is premature to think

> that the success rate of children at learning the structure-dependency of auxiliary fronting cannot be explained in terms of data-driven learning. The utterance tokens that could provide the crucial data apparently make up between 1% and 10% of interrogatives. A child obviously hears hundreds of thousands of sentences while engaged in language acquisition, and thus will hear thousands of examples that crucially confirm the structure-dependence of auxiliary fronting (p. 509).

With auxiliary fronting, he concludes, the "strongest and best-known pillar of support" for the poverty of the stimulus and hyperlearning collapses (ibid.).

The results of Pullum's corpus analysis can be challenged in several ways. First, it is debatable whether he adduces the right kind of positive evidence. His claim that questions of type (8) and (9) are relevant to the acquisition of complex polar interrogatives has not been investigated in developmental psychology or within computational learning models. Secondly, it is unquestionable that the examined corpus is not representative as a collection of linguistic material typically available to children.[6] To challenge Pullum's account, Legate and Yang (2002) looked at the frequency of the structures (8) and (9) in all files of the Nina corpus in the Childes database (MacWhinney, 2000). They found that the percentage of relevant sentences is actually much lower than Pullum and Scholz (2002) claimed. Although 44% of all sentences were questions, only 0.068% of these were of the (8) and (9) kind. Searching multiple Childes corpora, MacWhinney (2004) looked at the frequency of sentences of type (7) in conjunction with sentences containing auxiliaries in two positions, such as

(11) *Will the boy who is wearing a Yankee's cap step forward?*

He found only 1 such question in three million items. MacWhinney did not rigorously quantify the frequency of *wh*-questions with relative clauses (sentences of type (9)) in Childes. He claims, though, that "there are hundreds of input sentences of this type in the Childes corpus" (p. 890).

Thus, frequency estimates for relevant input vary considerably across the literature. The difference in several orders of magnitude results from different views on what counts as relevant evidence and which corpus is analyzed. What seems more important than the diversity of estimates, however, is the question how to test claims of relevance and frequency. Unfortunately, none of these authors comments on the empirical significance of the detected number of occurrences of input samples. Why would 1% of type (7)–(9) questions be sufficient to guide a child towards selecting the right kind of auxiliary fronting rule, but not, say, 0.001%? Is there a threshold frequency for learnability and why? Even if there was reliable data concerning the occurrence of various types of

[6] Pullum (1996) and Pullum and Scholz (2002) acknowledge the inadequacy of their source but of course some data is better than none. They concede that their corpus analysis is of preliminary nature and does not strictly refute arguments from the poverty of stimulus.

complex questions, this could merely undermine claims that there is *not any* evidence in child-directed speech. What is required in addition is a detailed account of how this evidence is processed and utilized in a concrete learning mechanism in order to determine whether there is *sufficient* evidence for purely data-driven learning to succeed.[7]

To summarize, it has been suggested by a number of authors (Sampson, 1989; Pullum, 1996; Pullum and Scholz, 2002; MacWhinney, 2004) that assumptions (ii) and (iii) behind the poverty of stimulus argument may be misguided. Perhaps complex polar interrogatives are sufficiently frequent in child-directed speech, perhaps other question types such as (8) and (9) are conducive to their acquisition, or even sufficient by themselves. Based on these suggestions I formulate

> Hypothesis-2: *Complex polar interrogatives can be learned from exposure to simple polar interrogatives, relative-clause constructions, and wh-questions with embeddings.*

which will be tested in the recursive Dual-path model in Section 7.3.2. This model provides an explicit learning mechanism for sentence production which may be capable of acquiring the syntax of question formation in the absence of positive evidence. Testing Hypothesis-2 in a computational framework, it might be possible to substantiate the idea that auxiliary fronting can be learned from simpler and similar structures whose occurrence is warranted in child-directed speech.

7.1.3 Statistical learning

Arguments against the data-driven learnability of complex yes/no-questions assume that positive examples of these structures are highly infrequent in the ambient language. While this might be the case, sparsity alone does not entail non-learnability unless it is also presupposed that children's learning mechanisms are not sensitive enough to exploit this evidence (assumption (ii)). In addition, such arguments often assume that by innate structure-dependent priming best explains the fact that children learn these constructions nonetheless (assumption (iv)). Recently, both these assumptions have been challenged by statistical approaches to language acquisition, in particular connectionist learning models such as simple-recurrent networks (SRN) (Elman, 1990, 1991). These models draw on the combined explanatory power of distributional properties of the input and domain-general learning mechanisms. Similar to an innately constrained language acquisition device, these models display processing biases qua architecture, but these biases are language-unspecific and adaptive. Constraints on processing are not

[7] In his keynote address to the 32nd Boston University Conference on Language Development 2007, O'Grady lamented that frequency haggling alone will not resolve the question whether some construction is learnable or not. For the case of 'want-to' contraction he sketched a learning account which does not rely on *any* positive evidence in the input. On this account, "the core properties of natural language syntax follow from the operation of an efficiency-driven [...] processor" (O'Grady et al., 2008). It remains open whether such an account could also be given for auxiliary fronting.

biologically fixed but can evolve in the course of learning and change with linguistic experience.

It has been demonstrated in many cognitive domains that SRN are useful for learning various types of statistical regularities in the input. They record the frequency of individual units of learning, e.g., lexical items, the frequency of co-occurrence of units, e.g., pairs of words, and the transitional probabilities between units, i.e., the predictive probability of one word, given the previous word. Accreting evidence from psycholinguistic research supports this approach to language learning in that children, even infants, are sensitive to statistical information in the input at various levels of language processing. For instance, Maye et al. (2002) have shown that infants' development of phoneme categories can be influenced by distinct distributions of speech sounds. Infants are also capable of segmenting continuous speech stream into words based on transitional probabilities in sequences of syllables (Saffran et al., 1996, Aslin et al., 1998, Saffran, 2001). Moreover, Gómez and Gerken (1999) have shown that infants can track adjacent and remote sequential dependencies in word ordering, thus acquiring the syntax of artificial finite-state grammars. An overview of many more recent results in statistical language learning can be found in Gómez (2007).

Hypotheses-1 & -2 conjecture that the syntax of complex polar questions could be assembled from more basic structures; simple yes/no-questions, relative clause constructions, and complex *wh*-questions. The process of structural generalization may be aided by the relative frequencies of substructures in grammatical and ungrammatical yes/no-questions in the input to a statistical learner. For example, grammatical yes/no-questions contain substructures such as

(12) ...*who* AUX VERB -ING or ...*who* AUX ADJECTIVE

whereas yes/no-questions which were formed by some structure-independent rule contain substructures such as

(13) ...*who* VERB -ING or ...*who* ADJECTIVE.

Substructures (13) from ungrammatical interrogatives, although possible substructures of other grammatical sentences,[8] may occur less frequently in a learner's linguistic environment than substructures (12) from grammatical interrogatives. If this is true for natural language input, and if the learner is sufficiently sensitive to substructure frequencies, she might form statistical expectations which facilitate the acquisition of correct polar interrogatives and interfere with the development of syntactic principles for ungrammatical polar interrogatives:

> Hypothesis-3: *Distributional information in the input is sufficiently rich for statistical learners to acquire the correct auxiliary fronting principle for yes/no-questions.*

[8] For instance, 'the athlete who diving killed' or 'the climber who hungry bears ate'.

This hypothesis has been tested in two previous studies involving SRNs, the work of Lewis and Elman (2001) and Reali and Christiansen (2005). These studies aimed at showing that when trained on English language input, SRNs are biased towards preferring grammatical over ungrammatical polar interrogatives in the absence of any sample of this construction in the learning environment. Before I describe the recursive Dual-path model approach to question learning, I will first review the results from these SRN studies.

The Lewis & Elman model

Lewis and Elman (2001) trained an SRN on an artificial English-like language containing simple polar questions and sentences with relative clauses. Then the network was tested on complex polar questions such as

(14) *Is the boy who is smoking crazy?*

At each position in the target sentence the SRN predicted a vector of lexical categories for the subsequent word in the sequence. The network's predictions for sentence (14) are depicted in Figure 7.1. After the initial segment *Is the boy...*, the SRN activated

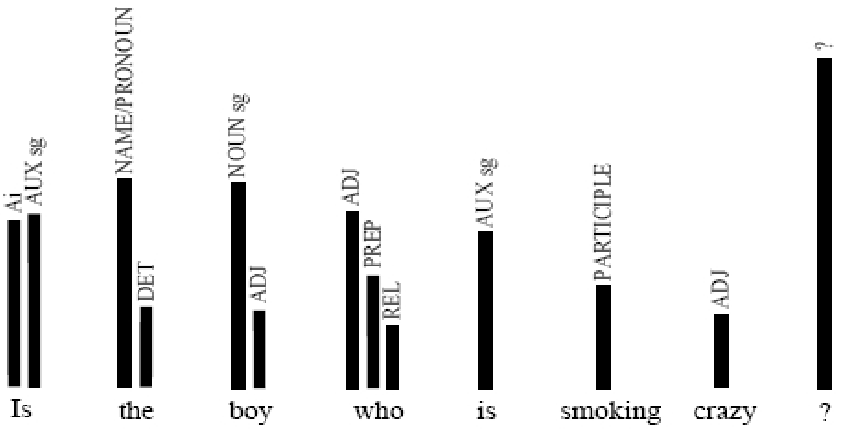


Figure 7.1: The SRN of Lewis and Elman (2001) tested on novel polar questions. The strength of the network's word category predictions is represented vertically above each target word.

the relative pronoun category at the position of *who* although the network had not experienced polar questions with embeddings in the input. Following the pronoun, the SRN activated the auxiliary category for the target *is* and did not activate a verb form as would be expected had the network learned structure-independent auxiliary fronting. Once the participle category for *smoking* was predicted, the network activated the adjective category at the relative clause boundary, even though these two word categories never co-occurred in training. Based on these results, Lewis and Elman (2001) suggest that "the network has formed an abstract representation of *aux*-questions, and generalized over the NP forms."

It is difficult to assess the significance of the Lewis and Elman (2001) results for a number of reasons. First, the authors provide too little information about their training set. It is not clear how much lexical and structural variation there was in the artificial language. One example of a declarative with relative clause is given, but it is not explained which types of relative clauses could occur in training (e.g., whether they could be subject- and object-modifier and relativize both roles). This example (*The boy who is smiling...*) matches the embedding of Figure 7.1 in that it is subject-relativized with a progressive verb form and this type of relative clause may have been the only one in the training language. It can also not be ruled out from the information given, that the language contained only embeddings attached to the first NP in the sentence. This would explain why the network activated a relative pronoun after *the boy* in the test structure. It seems premature to claim that the model has acquired a notion of relative clause or complex question formation on the basis of this input.

Secondly, Lewis and Elman report that in processing the sentence of Figure 7.1 the network did not predict a verb form after the pronoun *who*. This is taken as evidence for the correct structure-dependent rule. Presumably, however, we would expect activation of verb forms if the input language had contained subject-relatives with other verb tense/aspect (e.g., *...who smoked...*). Thus, it seems that the result of Figure 7.1 might be an artefact of the artificial language which may have been tailored to this specific learning problem. What's important is not so much what is in the language, but what may deliberately have been left out to strengthen the relevant substructure predictions. The network's preference for structure-dependent activation patterns might entirely be due to the absence of more diverse input.

Third, it seems that their model was only tested on a single complex polar interrogative so that the performance data of Figure 7.1 is anecdotal at best. Lewis and Elman (2001) do not report whether this result is even robust for different randomly generated training sets from the same language, or different random initializations of the network for the same training set. It is also unclear whether their model could handle different types of embeddings such as, e.g., object-relativized polar questions.

And finally, a more fundamental point of criticism of their approach. In the model's output profile of Figure 7.1 there is a 'path' of activation of correct categories that corresponds to a grammatical complex question. This path derives from the networks capability to *overlay* statistical expectations from substructures in different kinds of input sentences. This positive result by itself, however, does not show that their model has acquired the correct rule for complex question formation. Due to the nature of SRN mappings of words onto distributions of syntactic categories, there are many paths in the model's output profile which correspond to ungrammatical sentences as well. Take, for example, the sequence of most active categories in each sentence position of Figure 7.1, AUX PRONOUN NOUN ADJECTIVE AUX PARTICIPLE ADJECTIVE. This sequence does not form a segment of any grammatical English sentence. By parity of reasoning, the model has also 'learned' nonsense. Hence, it is not sufficient to handpick a path of activation which corresponds to the intended structure. Rather, it must be shown that there is no such path of activation for an ungrammatical complex polar question, or at least that such paths are less pronounced in terms of some reasonable error measure. In other words, it needs to be established quantitatively that after training the model is biased towards accepting grammatical polar interrogatives over ungrammatical ones.

The Reali & Christiansen model

Reali and Christiansen (2005) much improved on the Lewis and Elman (2001) approach to auxiliary fronting. They present two models of statistical learning which were trained on a more realistic, noisy input set, the Bernstein-Ratner corpus of mother-child interaction. This corpus did not contain any instance of a complex yes/no-question. The first model is the familiar n-gram model evaluated with cross-entropy (see Chapter 5). After extracting bigram and trigram frequencies from the corpus, sentence probability was computed for 100 matched pairs of grammatical and ungrammatical questions, e.g., the pair

(15) a. *Is the boy who is hungry nearby?*
b. **Is the boy who hungry is nearby?*

Comparing the cross-entropy for questions (15-a) and (15-b) indicates which one is more probable based on the distributional information in the training corpus. A question is classified correctly if the grammatical form has a lower cross-entropy than the ungrammatical form. Reali and Christiansen found that both the bigram and the trigram model correctly classified 96% of the tested questions. This behavior was stable for the actual questions that were elicited from children in the study of Crain and Nakayama (1987). Furthermore, classification was robust for both sets of questions when the n-gram models were trained on subsets of child-directed speech which corresponded to individual children subsumed in the Bernstein-Ratner corpus.

Since this corpus did not contain examples of complex polar interrogatives, the results of Reali and Christiansen (2005) indicate that there is sufficient *indirect* evidence in the primary linguistic data for children to learn the correct auxiliary fronting rule. Grammatical questions were favored by the n-gram models because the grammatical forms shared frequent word chunks with other sentences in the input. If children draw on similar kinds of statistical information in language acquisition, it might be unnecessary to postulate innate constraints on learning structure-dependent rules for question formation.

It has been shown that SRN are sensitive to bigram and trigram frequencies (see the references in Reali and Christiansen, 2005). Thus, it was a natural question to ask whether an SRN trained on the Bernstein-Ratner corpus would develop a comparable bias for structure-dependent auxiliary fronting as the n-gram models. Reali and Christiansen trained an SRN in this way and tested the network on 30 pairs of grammatical and ungrammatical questions such as (15-a)/(15-b). The trained network classified more than 83% of the grammatical structures (15-a) correctly as grammatical and misclassified less than 17% of the ungrammatical structures (15-b) as grammatical (Figure 7.2). In other words, after exposure to a corpus of child-directed speech the SRN was strongly biased towards preferring grammatical polar questions over ungrammatical ones. This suggests that domain-general learning mechanisms such as SRNs can reliably converge towards correct auxiliary fronting by attending to the rich distributional regularities in the learning environment. The more realistic natural language input to SRN in the Reali and Christiansen study guarantees that these results were not obtained by adjusting the training language to the specific learning task (e.g., by removing untoward constructions or skewing the distribution). Moreover, given the diversity of the corpus their results demonstrate that the SRN was able to exploit subtle statistical cues in the input to acquire auxiliary fronting.

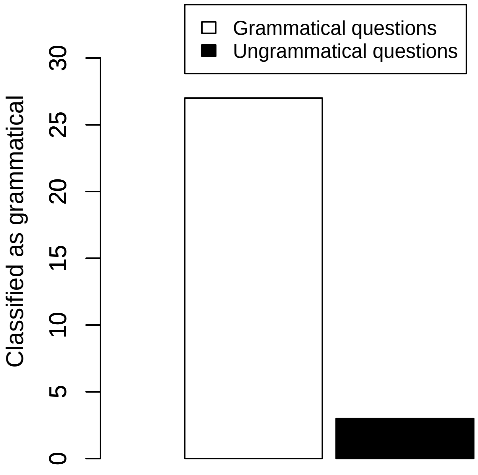


Figure 7.2: Grammaticality judgement of an SRN for polar questions; replicated from Reali and Christiansen (2005).

On the downside, Reali and Christiansen had to tag the input corpus with grammatical categories in order to train the SRN. This tagging removed a great deal of noisiness from the corpus. Children are not learning from tagged input but have to acquire syntactic categories from the ambient language during syntactic development. In general, the assumption of tagged input may not be exceedingly problematic since there is evidence that children are able to induce syntactic categories from distributional information alone (Gerken et al., 2005; Mintz et al., 2002; Redington et al., 1998). What seems slightly more problematic is the specific way the corpus was tagged in the experiment. Kam et al. (2007) note that a single tag PRON was used for all pronouns (interrogative, deictic and personal). This is relevant for learning auxiliary fronting because in order to correctly classify a test question such as *Is the man who...?* the model has to activate the initial sequence AUX DET NOUN PRON of word categories and then predict an auxiliary. To do this, the SRN must have developed statistical expectations that with some likelihood pronouns are followed by auxiliaries. With the pronoun category being so inclusive, the model can strengthen its expectations for the transition from pronouns to auxiliaries from sentences such as *He was hungry* or *What is going on?* because these sentences would have been tagged with a PRON AUX subsequence. Hence, this subsequence of word categories is supported by sentences which are functionally quite distinct from polar interrogatives. In successfully classifying yes/no-questions, the SRN may therefore have drawn on statistical information that is not available to children.

It also appears that auxiliaries were classified together with verbs into one category VERB. A subsequence of categories such as NOUN PRON VERB does not distinguish grammatical from ungrammatical complex polar questions, for instance, the pair

(16) a. *Is the boy who is watching TV hungry?*
b. **Is the boy who watching TV is hungry?*

The model's activation of the VERB category after the pronoun does not warrant the conclusion that it prefers the grammatical (16-a) over the ungrammatical form (16-b). Here, it is precisely the issue whether the model activates an auxiliary or a verb that determines whether it has acquired knowledge of correct auxiliary fronting. Putting auxiliaries and verbs into one category does not admit judgement in this case and the set of test sentences contained several sentence types such as (16-a), e.g.,

(17) *Is the dog that is sleeping on the blue bench?*

for which this point seems relevant. It is not clear, though, if and how this affected the results of Figure 7.2 because the model's performance was measured for the entire test questions, not just the initial segment up to the embedded auxiliary. Nonetheless, the word immediately following the pronoun would seem to be the crucial position to assess the model's behavior and this position is ambiguous due to the particular tagging. This suggests that the model's good classification performance may to some extent depend on the categorization used to tag the training corpus. It should be emphasized again, however, that the main results of the Reali and Christiansen (2005) paper were obtained with an n-gram model which is more general than the SRN. The positive results on the data-driven learnability of complex polar questions which were obtained with this model did not in any way depend on the assumptions made in tagging.

Reali and Christiansen tested the SRN on subject-relativized polar interrogatives (with one object-relativized exception which was classified correctly) and all tested structures contained the auxiliary *is* in both the main and the embedded clause. It remains to be determined whether the model's behavior is robust with respect to more structural variation in the test items. The distributional regularities on which the SRN draws to judge the grammaticality of test items such as (15-a)/(15-b) may not be sufficient to classify other types of complex questions.

7.2 The recursive Dual-path model approach

In my own approach, I aimed at extending the results of Lewis and Elman (2001) and Reali and Christiansen (2005) using the recursive Dual-path model. This model contains an SRN as a sequencing subnetwork and should therefore be similarly sensitive to distributional regularities in the input. The experimental work presented here, however, differs in important respects from the SRN studies. In these studies, the SRNs mapped input sequences onto sequences of word categories and at each sentence position they activated a distribution of categories which could be interpreted as possible grammatical continuations according to the networks' experience. Whether a particular structure was learned, and how well, depended on how strong the activation of the correct word category at each position was, compared with activations for competing structures. In order to measure the model's performance, it was therefore subjected to a grammaticality judgement task for correct and incorrect polar question, and the resulting errors were compared to determine whether the SRN was biased towards the grammatical or ungrammatical sentence form. The recursive Dual-path model, on the other hand, maps meaning representations onto sequences of lexical items. At each sentence position only the most active lexical item is recorded as output and this property allows us to compare actual productions word-by-word against intended productions. The model is given an input message which it has to cast into a lexically definite sentence. Because this input message is neutral between correct and incorrect polar forms, we can directly measure syntactic choices in the model's productions. Hence, the recursive Dual-path model approach replaces grammaticality judgement with testing active syntactic knowledge required to correctly produce a specific complex polar question from a given message. This makes the model's learning task considerably more difficult than the SRN's task. In order to produce a correct question, it is not sufficient that the model shows some activation of the correct lexical item at each sentence position, the level of activation must be higher than that of any other word competing for this slot. Secondly, there was more variation in the output range of the recursive Dual-path model because the lexicon of the language used to train the model contained more items than there were word categories in the Reali and Christiansen (2005) experiments. And third, the model learned from untagged input and had to induce syntactic categories by itself, based on co-occurrence frequencies of words in sentences of the training set.

On the other hand, the input to the recursive Dual-path model was generated from an artificial grammar. Thus, it lacked the structural variation, noisiness and more realistic distribution of the Bernstein-Ratner corpus employed by Reali and Christiansen. Moreover, in contrast to the SRN approach, I assumed that children use meaning in syntax acquisition. The recursive Dual-path model was learning the target language from exposure to sentences paired with their meaning, not from sentence input alone, and these meaning representations played a critical role in the generalization to novel utterances. Semantic similarities between messages for experienced and novel constructions enabled the model to generalize to novel constructions which were not experienced during learning. My approach also differed from the SRN experiments in that I tested the model on a wide range of complex yes/no-questions with intransitive, active/passive transitive, oblique, prepositional dative and ditransitive constructions occurring as main and embedded clauses in all combinations. In addition, main and embedded clause auxiliaries were allowed to differ, and any grammatical role could be relativized inside the embedded clause. Thus, tested utterances could be object- as well as subject-relativized. More variation in the test items ensured that positive results were less likely to be contingent on the specific learning conditions, and therefore more robust and relevant to the poverty of stimulus debate.

In the experiments, I focused on identifying an explicit learning mechanism for polar interrogatives in production, not merely a bias for grammaticality. This mechanism drew on distributional regularities in the input, such as frequent substructures, but more importantly it relied on the presence of particular constructions from which the model generalized to novel constructions in a combinatory fashion because of semantic similarities in the message. I sought to isolate the input factors that contribute to this goal by systematically varying the properties of the training language. In this way I can show that the model is able to learn polar interrogatives in some conditions but not in others and this will help to explain why generalization occurred. I will argue that the model's behavior is not depending solely on n-gram statistical regularities in the input but that it is acquiring structure-dependent knowledge of complex question formation which can not be reduced to transitional probabilities in frequent substructures. Furthermore, I attempt to trace the source of generalization by analyzing the HIDDEN-layer representations the model developed during learning and I will argue that clausal integrity which prohibits auxiliary extraction from embeddings transfers to complex polar questions by analogy from similar structures in the input.

7.2.1 Artificial language, semantics and method

The recursive Dual-path model was trained on untagged input from a structurally rich artificial, English-like language which allowed the creation of simple clause sentences, various relative clause structures, and different types of questions. Table 7.1 provides an overview of the basic constructions in the language from which more complex sentences were assembled. Relative-clause structures in the language contained a main clause and one subordinate clause drawn from these structures. Every combination of simple clause types was permitted and relative clauses could be attached to any NP in the main clause, regardless of its grammatical function. Hence, there were center-embedded and right-branching sentences in the language and the head noun could occupy any grammatical role in the relative clause which met its animacy constraints. Moreover, the language allowed the formation of three classes of questions, simple polar questions (SPQ from now on), complex polar questions (CPQ from now on), and complex *wh*-questions; Table 7.2 lists examples of these structures. The input grammar had verb tense and aspect, inflectional morphemes were represented as separate lexical items. As others have argued (Pullum and Scholz, 2002; MacWhinney, 2004), I suggest that the syntax of complex yes/no-questions can be assembled piecemeal from simpler and similar constructions

Structure	Example
Intransitive	`the cat was sleep -ing .`
Transitive	`the woman kick -ed the teacher .`
Transitive passive	`the teacher was kick -ed by the woman .`
Prepositional dative	`a girl throw -s the stick to the cat .`
Double object dative	`a girl throw -s the cat the stick .`
Oblique	`the nurse is play -ing with a dog .`

Table 7.1: Basic construction types in the language to which the recursive Dual-path model was exposed.

Question Type	Example
Simple polar question	`is the cat chase -ing the dog ?`
Complex polar question	`was the cat that is chase -ing the dog run -ing ?`
Complex *wh*-question	`who is the cat that is chase -ing the dog run -ing with ?`

Table 7.2: Classes of questions admissible in the artificial language.

which are warranted in a child's linguistic environment. For example, subject-auxiliary inversion might be learned from simple yes/no-questions in the input, and auxiliary extraction across a relative clause might be learned from complex *wh*-questions.

To examine whether the model acquires a structure-dependent or structure-independent principle of question formation, it is sufficient to employ a language with center-embedded yes/no-questions and progressive aspect in each clause. Nonetheless, my language allowed complex questions to be created in which any grammatical role could be modified (or relativized, for that matter), just as in the declarative structures. In addition, verb tense and aspect of each clause could form any combination within a complex question. For instance, I also tested the model on questions such as `was the cat chase -ing the dog that play -ed with the girl ?`, a right-branching yes/no-question with simple aspect in the relative clause, since this enabled me to analyze the model's behavior in more detail across the different input conditions.

The artificial language used to train the model contained distinct 271 construction frames. The lexicon comprised 57 words in 15 categories which filled appropriate slots in these frames. This allowed the creation of roughly 11.3 × 10^9 different sentences in this language.[9] The model was trained on a set of 10.000 sentences, which were randomly generated from this language, for a total of 100.000 epochs.[10] To test the model, questions with relative clauses were generated. The language allowed the creation of roughly 5.3 × 10^9 different CPQ. 1.000 questions were randomly selected and the model was tested periodically on these items after every 5.000 training sentences.

Sentences generated for training and testing were paired with their corresponding meaning, and the model was exposed to these pairs as described in Chapter 3. To represent the semantics of relative clauses, I used the SIMPLE TOPIC-FOCUS message with *gap-link* which was discussed in Chapter 4. The message had event alternation, so that the order of events (and therefore clauses) in complex sentences was not encoded spatially. The only way the model could determine which message features corresponded to which clause (e.g., in complex questions) was by attending to the topic/focus features which marked the head noun and its semantic role in the relative clause. In other words, there was no default order of events in the semantic representations, features in the message were not clause-level specific. Questions were distinguished from declaratives by

[9] Because training conditions differed in the experiments described subsequently, the expressivity of the input language also varied slightly across conditions.

[10] One epoch corresponded to one sentence in training.

using question features in the message. There were two types of question features, one feature for polar questions, and several features for *wh*-questions. The message for SPQ was identical with the message of the corresponding declarative, except that in addition a question feature signalled to the model that a different ordering of constituents was intended. The message for complex *wh*-questions was slightly more complicated. Consider the *wh*-question

(18) `who was a cat that the dog is give -ing a toy to walk -ing with ?`

In order for the model to produce such questions correctly, the sentence message needed to represent which main clause constituent was co-referential with the interrogative pronoun `who`. Suppose it was co-referential with `the girl` in the declarative

(19) `a cat that the dog is give -ing a toy to was walk -ing with a girl .`

Then the message for question (18) would be identical to the message for the declarative (19) except that in addition a *wh*-feature marked the role of `the girl` in the event semantics. This feature would indicate that a question, not a declarative, was intended and signal to the model that `the girl` had to be omitted from the surface form. And finally, the message for CPQ such as

(20) `was a girl that throw -s a cat the apple being push -par by a boy ?`

was identical to the message for the corresponding declarative

(21) `a girl that throw -s a cat the apple was being push -par by a boy .`

plus the polar question feature which was also used to distinguish SPQ from single-clause declaratives.

It is important to point out here that polar question features, unlike all other features in the message, could not alternate between events. Polar question features were rigid and not associated with a particular atomic event within a complex proposition. Since events could alternate in the message, but question features could not, the model was not aided towards associating an active question feature with a particular clause in the sentence. In other words, in testing the model on CPQ, it received a message for a complex declarative plus a *clause-neutral* question feature. In this way, the model was unbiased with respect to which auxiliary (main or embedded) it had to produce at sentence onset. This was crucial for testing whether the model favored a structure-dependent over a structure-independent principle for auxiliary fronting.

7.3 Modelling results

Arguments against the poverty of the stimulus are often formulated as verbal theories of acquisition which are difficult to test because no explicit learning mechanism is provided. The recursive Dual-path model provides a such a mechanism which has been shown to explain several aspects of syntactic development (Chang, Dell, and Bock, 2006; Fitz and Chang, 2008). Using this model, I wanted to identify input conditions which are sufficient to acquire the syntax of CPQ in the absence of positive exemplars in the input. With the described language and question semantics I tested the hypotheses of Section 7.1 in a series of simulations. In these experiments, every parameter of the model was kept constant, for instance network size, random initialization, and the learning mechanism itself. Likewise, the model was tested on the same set of randomly generated CPQ in all conditions. The only variable across experiments was the kind of input structures the model received in training. In this way I could determine the effect of the input on the learnability of CPQ. This allowed me to assess the validity of the formulated hypotheses in a methodologically rigorous manner and to substantiate some of the claims which have been made in the literature against the supposed non-learnability of polar questions. The experiments will be described in the following sections. All reported results were averaged over ten model subjects, which differed in terms of the randomly generated training items to which they were exposed in each condition.

7.3.1 Simple polar interrogatives

I first looked at Hypothesis-1, viz the claim that CPQ can be learned from simple polar questions and relative-clause constructions in the absence of positive evidence. To test this hypothesis, the model was trained on 10.000 sentences from the artificial language, with a simple- to relative-clause ratio of 1, and roughly 15% of all training items being SPQ. The distribution was not intended to match human linguistic input. It was required that the model learned all input structures to an adult degree at the end of training. The model was tested on the critical structures, polar interrogatives with a center-embedded relative clause, but also on right-branching structures. I will refer to these structures as CPQCE and CPQRB, respectively. In this condition, the model produced no correct CPQ when measured in terms of *sentence accuracy* (perfect match) or *grammaticality*. I therefore used the *production error* measure as defined in Chapter 6 to evaluate the model's performance. Production error is a string-distance measure scaled by the length of the target utterance. Unlike sentence accuracy, production error is sensitive to partially correct productions and to misplaced chunks of correct constituents. Informally, production error yields the percentage of lexical errors the model made, out of the maximal number of errors that the model could have made in a given utterance. The results of testing the model on CPQCE and CPQRB at an adult state are shown in Figure 7.3. The model's performance is compared with an input condition in which the training set contained only declaratives and no questions of any kind. It is apparent that the model performed much better on CPQRB than on CPQCE and this difference was statistically significant in both conditions (no questions: $F(1,9) = 139.9$, $p < 0.001$; simple polar questions: $F(1,9) = 77.1$, $p < 0.001$).[11] This behavior for novel interrogatives is in line with the observations

[11] Repeated measures ANOVA at the end of training.

made in Chapter 6 regarding declaratives, where it was shown that the model learned right-branching faster than center-embedded constructions for different levels of embedding. Secondly, the error level on CPQCE was equally high in both conditions, with and without SPQ in the input; there was no statistical difference detectable (F(1,9) = 1.2, p = 0.3). This suggests that the model was not aided in learning auxiliary fronting for CPQCE by being exposed to SPQ. This indicates that Hypothesis-1 might be false in the recursive Dual-path model framework. Relative clause constructions and simple yes/no-questions are not sufficient for the model to assemble the syntax of CPQCE. However, it can be observed that SPQ seem to aid the acquisition of CPQRB. The error level on these structures dropped significantly between the two input conditions (F(1,9) = 11.4, p < 0.01). In CPQRB, the embedded clause follows the main clause. Thus it is the first auxiliary of the declarative which is placed in front when such a question is formed, just as in SPQ. The results therefore seem to suggest that SPQ in the input might lead the model to entertain a *structure-independent* rather than a structure-dependent hypothesis about complex question formation. Not only do SPQ not help in the acquisition of CPQCE, they might even be detrimental to learning these structures because they seem to bias the model towards adopting the wrong auxiliary fronting principle.

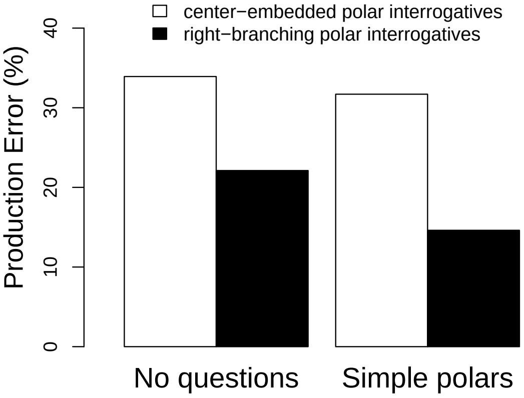


Figure 7.3: Mean production error for CPQCE and CPQRB with simple polar questions in the input.

7.3.2 Complex *wh*-questions

In the next condition, I examined Hypothesis-2, the claim that CPQ can be learned from simple polar interrogatives, relative-clause constructions and *wh*-questions with embeddings. Here, the idea is that a learner might be able to recruit knowledge about the syntax of CPQ from the syntax of various other constructions with which these questions share surface similarities. Simple questions contribute knowledge about subject-auxiliary inversion, knowledge of relative clause formation in declaratives transfers to relative clauses in complex questions, and from complex *wh*-questions it might be inferred that main clause auxiliaries can be fronted across a relative clause. I tested this synthetic approach to CPQ learning in the recursive Dual-path model. As in the previous experiment, the input language contained roughly 50% sentences with relative clauses and 15% SPQ. Around 10% of all training items were complex *wh*-questions, either with a center-embedded or a right-branching relative clause. Again, the distribution was designed to ensure that the model acquired all input structures with more than 90% accuracy (in terms of perfect match). The model was tested on 500 CPQCE and CPQRB, respectively. In this condition, we observe a performance pattern complementary to the previous conditions (Figure 7.4; for comparison, the data of Figure 7.3 was reproduced). Compared to these conditions, the production error for CPQCE dropped significantly when the model was exposed to complex *wh*-questions in training (F(1,9) = 19.9, p < 0.05). However, the error level of CPQRB remained constant (F(1,9) = 0.01, p = 0.9) between the 'simple polars' and '*wh*-questions' conditions. Yet, the model still performed superior on CPQRB than on CPQCE, although the difference was not statistically significant (F(1,9) = 3.9, p = 0.07). Because both these structures were entirely absent from the input, this result adds to the robustness of the model's preference for right-branching over center-embedded structures. The constant error level of CPQRB suggests that the learnability of these structures did not profit from the complex *wh*-questions in the environment. Syntactic knowledge of auxiliary fronting in CPQRB that the model needs to extract from the input is already contained in relative clause declaratives and SPQ. Exposure to right-branching *wh*-questions did not seem to add information relevant for the acquisition of these structures. The model's performance on CPQCE, on the other hand, improved significantly compared to both conditions without complex *wh*-questions. These questions aided the process of correctly sequencing CPQCE, as is witnessed by a lower overall production error. This suggests that Hypothesis-2 is supported within the framework of the recursive Dual-path model.

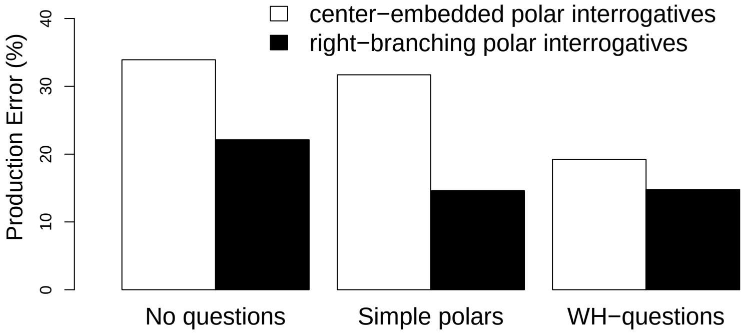


Figure 7.4: Mean production error for CPQCE and CPQRB with complex *wh*-questions in the input.

It remains to be determined through analysis, however, what the nature and degree of support for this hypothesis is. The fact that production error for the entire CPQCE test items dropped considerably only indicates that the '*wh*-questions' condition is conducive to learning these structures. The scores do not reveal what kind of errors the model made and whether it actually produced any correct CPQCE structures. Moreover, they do not tell us in what sense CPQCE can be learned in this input condition, and whether the model does indeed favor a structure-dependent over a structure-independent principle of CPQCE-formation. These issues will be addressed in the remainder of this chapter.

Error analysis

I examined 100 CPQCE with progressive aspect or passive voice in both clauses, which the model produced at the end of training with *wh*-questions in the input.[12] The dominant error types classified into four categories. Type I errors involved the repetition of the main clause auxiliary after the relative clause was complete, e.g., **Is a dog that was being kicked by a sister is giving the boy the tomato?* Crain and Nakayama (1987) refer to this type of error as 'prefixing'. When the target main clause was produced correctly but there was any kind of scrambling in the relative clause, I call this pattern Type II error (with the exception described as Type III next). A Type III error occurred when the model formed CPQCE in accordance with a structure-independent hypothesis and placed the relative clause auxiliary in front, for example in **Is a woman that a brother walking with is throwing a man a paper?* Occasionally, the model set out to produce a *wh*-question instead of a yes/no-question by replacing the sentence initial auxiliary with the interrogative pronoun *who.* I refer to this behavior as Type IV error. Table 7.3 summarizes the distribution of error types found in the model's output. Mixed errors occurred in conjunction with other errors, genuine errors occurred as sole errors within one test sentence. On both scales Type I errors were the most frequent. This indicates that the model had difficulty extracting the main clause auxiliary after the embedded clause, the sentence-initial auxiliary was merely a duplicate. Type II errors show that relative clause integrity was difficult to maintain for the model when relative clauses were combined with main clause polar questions into novel complex constructions. Nonetheless, for the majority of test sentences the model produced a correct embedding despite not having been exposed to any such construction in training. This suggests that there was substantial structural transfer from declarative sentences with embedding (which the model learned to perfection) to interrogatives with embedding. Genuine Type III errors did not occur in the '*wh*-questions' condition. In other words, the model did not displace the embedded clause auxiliary. Such errors would have indicated that the model entertained a structure-independent rule of complex question formation. The absence of these errors, however, does not entail that the model was inclined towards the *correct* principle. To establish this, a more detailed analysis of Type I errors was required (see below). The low Type IV error rate indicates that the model did not confuse *wh*- and yes/no-questions when given a message to produce a complex interrogative. Hence, the decrease in production error in the *wh*-condition of Figure 7.4

	Type I	Type II	Type III	Type IV	Other	Correct
Mixed	48	47	4	1	19	8
Genuine	27	21	0	0	17	8

Table 7.3: Types of errors for 100 CPQCE test items in the trained model.

[12] I picked the model subject which was closest to mean performance in terms of sentence accuracy and looked at the first 100 items in the test set.

was not due to the model producing structurally similar *wh*-questions when tested on CPQCE messages, but reflects a genuine improvement of yes/no-question learning.

In the Crain and Nakayama (1987) study Type I errors accounted for 58% of all errors and were the most frequent errors children made (a very similar Type I error rate was recently reported in Ambridge et al., 2008.). Type III errors did not occur in their data. Mixed (genuine) Type I errors accounted for 40% (42%) of all errors and were the most frequent errors in my experiment, and since in addition genuine Type III errors were absent, the model's error profile matched the developmental data quite well.[13] Crain and Nakayama argue that the types of errors they found support the claim that children have knowledge of the correct structure-dependent rule for auxiliary fronting. They acknowledge, however, that the high frequency of Type I errors puts this claim in jeopardy. These errors could be the result of auxiliary duplication in the main or embedded clause. If they derive from the latter, this would rather support the idea that children initially entertain a structure-independent rule and retract from it later in development. Thus, Crain and Nakayama needed to show that Type I errors resulted from duplicating the main clause auxiliary and did not reflect a lack of syntactic competence. They conducted a second experiment in which they exchanged the embedded clause auxiliary with a modal verb, e.g., *Is the boy who can see Mickey Mouse happy?*, the procedure remained the same. In this way, they suggest, it can be determined whether children's 'prefixing' errors resulted from fronting the embedded clause auxiliary. Because no tested child produced questions such as **Can the boy who can see Mickey Mouse is happy?* they concluded that in both experiments Type I errors were not due to copying the leftmost auxiliary. Therefore, they argued, 'prefixing' errors were not invalidating the claim that children have the correct syntactic knowledge for complex question formation.

It is questionable, however, whether this experiment is methodologically sound because it removes the ambiguity between main/embedded clause auxiliaries in the experimenter's instruction for the task. One of the major advantages of computational models over developmental studies with children is that we can analyze the internal representations the model has developed in the course of syntactic development. Inspecting these representations may allow us to determine the origin of Type I errors in the model—whether they result from prefixing the wrong auxiliary or not. For all test sentences which led to genuine Type I errors in Table 7.3 I recorded the activation of units at the WHERE-layer of the model as it predicted the lexical items in these sentences. Recall that these sentences were produced correctly up until the main clause verb in which position a third auxiliary was erroneously inserted. I quantized activation levels into five distinct states and plotted the state vectors of the WHERE-layer units (0A, 0X, 0Y,..., 1Z) against the word category of the produced constituents. Figure 7.5 displays such a

[13] Crain and Nakayama's Type II errors involved 'restarting' after the completed relative clause, e.g., **Is the boy that is watching Mickey Mouse, is he happy?* My artificial language did not have personal pronouns, so these errors were not discernable. Type II errors in the model did not occur in their study. This may be due to the fact that their methodology involved the repetition of a correct relative clause whereas the modelling experiment resembles elicited production.

sequence of activation vectors for the question `is a cat that the ball is push -par`

Category	WHERE-layer (thematic roles)							
	0A	0X	0Y	0Z	1A	1X	1Y	1Z
AUX			■		■	■		
DET	░		▒		■	■		
NOUN			■			■		
THAT	░				░			
DET	▒		■		■			
NOUN			■					
AUX	■							
VERB	■							
PREP	■		▒	▒	■			
VERB			░	▒	■			

Figure 7.5: WHERE-layer activation for the prefixing error.

`by chase -ing the dog ?` Dark cells indicate strong activation, white cells represent inactive units. WHERE-layer units corresponded to thematic roles and in acquiring the target language, the model learned to sequence these units in order to express the input message. At the beginning of this question, the model strongly activated the action unit (1A) and the transitive agent unit (1X) of the main clause. Furthermore, it activated the transitive patient unit (0Y) of the embedded clause. This pattern of activation indicates that the model was in a state of uncertainty about the sentence structure to be produced. First, there was uncertainty about which was the main and which the embedded clause of the target sentence. Unit 1X would be the right choice if the declarative `a cat that the ball...` was to be produced, unit 1Y if the declarative `the ball is push -par by...` was to be produced. Secondly, there was uncertainty whether the target sentence was a declarative or an interrogative as indicated by the active action unit 1A. This unit carried information about the tense and aspect of the main clause verb and needed to be activated to produce the sentence-initial auxiliary in an interrogative. Thus, there was strong competition between units at the sentence onset. Since the model correctly produced an interrogative, the unit 1A won the competition in this case. The model proceeded with the main clause agent `a cat`, the pronoun `that`, and so forth. The crucial aspect of this plot is that at the sentence-initial position the model displayed zero activation of the embedded clause action unit 0A. This demonstrates that the first auxiliary of the target polar question derived from the main clause verb, not the embedded clause verb. In other words, the model was in a state of complete certainty that the embedded clause auxiliary was not the correct constituent to initialize the production of the intended structure. Hence, the prefixing error in the model did not result from duplicating the wrong (embedded clause) auxiliary.[14] It can also be observed in Figure 7.5 that the model activated the embedded clause action unit 0A—and only this unit—at the position of the embedded clause auxiliary. At the position where the prefixing error was manifest in the surface form (bottom row of Figure 7.5) the model again activated the main clause action unit 1A and there was no sign of activation at the embedded clause action unit. Both observations confirm the model's *structure-dependent* approach to the formation of CPQCE in the absence of positive evidence in the input.

So far I looked at characteristic errors the model produced in testing, but Table 7.3 also shows that it generated correct CPQCE structures (rightmost column). Correctness is measured here in terms of perfect match, i.e., the model produced the target structure exactly, with 0% production error. Thus, the model spontaneously generated correct CPQCE in the '*wh*-questions' condition, i.e., without having been exposed to any such question in the learning phase. Averaged over all subjects the accuracy of CPQCE at the end of training was around 7%, CPQRB reached around 16%.[15] Examples of questions which the model produced correctly are listed in Table 7.4. Note that embeddings in the examples of correct CPQCE comprise subject- as well as object-relatives, actives and passives, and transitives and datives. Thus the model's capacity to generalize was not restricted to a narrow set of constructions, or even a single type of polar question. Furthermore, correct productions included different auxiliaries (*is* and *was*) in the main

Correct complex polar questions
Center-embedded
`is the boy that is being hit -par by the cat give -ing a truck to the father ?`
`was a father that was show -ing the sister a beer throw -ing a phone to a man ?`
`is the dog that a cat was present -ing a apple show -ing the brother a milk ?`
Right-branching
`is a girl hit -ing a mother that is walk -ing ?`
`is a girl run -ing with a sister that the man is jump -ing with ?`
`is a mother walk -ing with a man that was being push -par by the woman ?`

Table 7.4: Examples of yes/no-questions with relative clause which the model produced correctly when exposed to *wh*-questions in training.

[14] WHERE-layer activation states were inspected for all other test questions in which Type I errors occurred and showed a similar pattern. In all cases, the main clause action unit was fully active while there never was more than the faintest activation of the embedded clause action unit.

[15] Confer the Appendix B (page 287) to this chapter for results which significantly improved on this accuracy.

and embedded clause. This indicates that generalization was not contingent on the construction type of the tested items which adds to the robustness of the model's behavior.

Through a detailed error analysis I established three interesting properties of the model's performance in the '*wh*-questions' condition. First, the error profile qualitatively matched the types and frequencies of errors children made in the Crain and Nakayama (1987) experiments. This suggests that the model implements learning mechanisms which are adequate from a psycholinguistic point of view to study the syntactic development of auxiliary fronting. Secondly, despite producing many errors overall, the model was strongly biased towards structure-dependent question formation. This was witnessed by the fact that prefixing errors were exclusively due to faulty reactivation of the main clause auxiliary at the verb position plus the absence of errors that involved extraction of the embedded clause auxiliary. And third, the model managed to actually produce a fair amount of correct CPQ without exposure to positive examples. Crain and Nakayama aimed at showing that in the acquisition of complex questions, children do not retract from false syntactic principles. Rather, they entertain the correct structure-dependent hypothesis already early in syntactic development although competing hypotheses are simpler, consistent with, and warranted by their linguistic experience. They argued that these findings can only be explained by assuming that structure-dependence is an innately endowed principle of universal grammar. In my simulations, I found that the lack of positive examples of CPQ in the input did not mislead the model into adopting an ungrammatical, structure-independent principle either. To explain this behavior, however, it is not necessary to stipulate a traditional notion of universal grammar. The model achieved this by recruiting pieces of information from different input structures, combining them into knowledge of auxiliary fronting in polar questions. From simple polar questions the model learned sentence-initial subject/auxiliary inversion. From complex declaratives it learned how to modify noun phrases, to suppress the head noun in the relative clause, and acquired a notion of clausal integrity. And from complex *wh*-questions it learned the grammaticality of subject/auxiliary inversion across embeddings. In this manner, the model developed dispositions to generate CPQ, the correctness of which was manifest in the error profile, as well as the spontaneous, error-free production of such questions. Because the structures that the model drew on to assemble the syntax of CPQ are warranted in child-directed speech, the model provides an alternative, data-driven explanation for the learnability of these constructions.

7.3.3 Bootstrapping

In this section I will describe a training regime designed to strengthen polar interrogative learning in the model. In this regime, the model received its own spontaneous correct productions of complex yes/no-questions as input in the subsequent learning cycle. Since this procedure makes rather controversial assumptions about a learner's linguistic environment, I will first attempt to motivate it and then discuss the results. The crucial message of this chapter is that structure-dependent auxiliary fronting in interrogatives can be learned from simpler and similar structures in the linguistic input.

The previous '*wh*-condition' provided a proof of concept for this idea although overall sentence accuracy for CPQCE remained low.[16] It should therefore be pointed out, that the main results of this chapter do not hinge on the current section. This bootstrapping section was intended to pinpoint some of the disadvantages the model might have compared with children in that the standard learning procedure might be too rigid, and to suggest some means of modelling a more realistic learning environment.

The guiding objective of a human language learner is to be able to engage in meaningful and successful communication. The acquisition of syntax subserves this objective, but it is not the primary learning target. An important aspect of human communication is the ability to convey one's intentions and to influence and manipulate the intentions of others (Tomasello, 2003). On this view, language is a cultural skill to communicate intentions and desires, to initiate or prohibit actions, and to steer the attention of others to objects or events in a joint frame of reference, rather than a conventionalized system to exchange facts about the world. Thus, language use is essentially goal-directed, linguistic utterances aim at altering the mental state of others. This intentional activity requires taking into account the intentions of others, and this is a reciprocal relation between speaker and hearer. Language use therefore is interactive in a non-trivial sense. Without an understanding of others as intentional agents there is no meaningful communication. Moreover, due to the goal-directed nature of language use, there is a value attached to communicative success. To achieve the goal of changing other people's mental states through linguistic utterances, language systems must be sufficiently aligned. The closer they are aligned the higher the chances to achieve a communicative goal. In discourse, speaker and hearer will strive to increase their communicative success. For brevity, let's say that discourse participants attempt to maximize (the realization of) some pragmatic function in communication (such as conveying one's needs or intentions, re-directing attention, and influencing mental states of others.).

In the context of child language acquisition, these general remarks may have important implications. Some syntactic constructions might be difficult to learn because they are not as useful as other constructions in terms of their pragmatic function. For instance, complex polar questions are useful for children to express wishes and requests when addressing their parents. Referring to the child's own needs and desires, such requests are likely to take a first person pronominal subject in the main clause, for example

(22) *Can I have the cookie that's on the table?*

Relative clauses, however, rarely modify pronominal NPs; therefore center-embedded yes/no-questions with a full subject NP such as

(23) *Can the dog that's sleeping come with us?*

might be less pragmatically useful and hence less frequently used than right-branching

[16] Again, confer the Appendix B (page 287) to this chapter for stronger results in this condition.

polar questions. There is ample evidence that frequency of occurrence affects sentence comprehension, production and language acquisition in general, because high frequency strengthens the representations of linguistic constructions in memory which facilitates their activation in language use (Diessel, 2007).

Yet, there is also a learning pressure to be precise in communication. The more accurately a human learner is able to express his/her communicative intentions (according to the social environment's linguistic standards) the higher the chances that the pragmatic function of speech will be realized (e.g., the satisfaction of a child's needs and wishes). This might motivate a child to acquire constructions such as (23) and use them when they are appropriate to express some communicative intention. Once this construction is produced correctly and associated with some specific pragmatic function, it might then become entrenched in grammar through more frequent use in the learner's speech. Conversely, a construction mastered by the child increases the options for successful communication for the parents. This might lead parents to use this construction more frequently when pragmatically felicitous. A child might profit in turn from this dynamic change in the distribution of parental speech in that it helps to consolidate the child's memory of the underlying syntax. Thus, language learning may be mediated by performance-related parental feedback following grammatical utterances that achieve their communicative function. This kind of 'covert positive reinforcement' is not to be confused with explicit feedback, e.g., parents overtly correcting children by telling them the grammatical utterances they should have used.

There is some evidence that corrective parental feedback is available and effective in guiding the semantics of word learning (Brown and Hanlon, 1970) and that the acquisition of foreign language vocabulary is a function of differential reinforcement (Whitehurst and Valdez-Menchaca, 1988). The effect of social feedback on syntactic development is less well understood and particularly controversial (Bohannon and Stanowicz, 1988; Bohannon et al., 1990; Brown and Hanlon, 1970; Demetras et al., 1986; Hirsh-Pasek et al., 1984; Moerk, 1991; Morgan and Travis, 1989; Morgan et al., 1995; Nelson et al., 1984; Penner, 1987). Several of these studies, however, indicate that children's grammatical utterances are more likely to elicit exact parental repetitions than ungrammatical utterances (Bohannon and Stanowicz, 1988; Demetras et al., 1986; Hirsh-Pasek et al., 1984; Morgan and Travis, 1989). Bohannon and Stanowicz (1988), for instance, found in 16 transcribed child-parent conversations that parents were differentially responsive to the grammaticality of children's speech. Exact parental repetitions occurred after 10-12% of all child utterances and followed almost exclusively children's *grammatical* utterances (90%). This data suggests that children are exposed to more grammatical positives of some construction type in response to their own correct use. Such parental feedback can be instantaneous, as attested by these studies, or it might be delayed and not result from conscious parental effort or awareness. This kind of covert positive reinforcement through repetition might skew the distribution of child-directed speech towards those constructions which have been uttered spontaneously by the child and this may provide a child with the frequency of occurrence necessary to consolidate the underlying syntax in memory. Moreover, Penner (1987) and Demetras et al. (1986) found that parents were far more likely to continue or expand the current topic of discourse following grammatical utterances than after ungrammatical utterances. In other words, parents were differentially responding to grammaticality in that children's grammatical utterances caused parents to move on in conversation whereas ungrammatical utterances tended to disrupt the flow of conversation. This tendency for topic extension to occur more frequently after grammatical utterances might provide another subtle but useful cue to children in language acquisition. When child-parent conversation is uninterrupted and a grammatical utterance is followed by parental extension of the topic, the child receives covert positive information that her utterance was comprehended and achieved its communicative function. This kind of parental response might create a selective pressure which encourages and increases children's active use of more difficult and mature grammatical constructions.

In sum, several studies of child-parent interaction suggest that various types of positive parental feedback succeed grammatical utterances of children in natural discourse. Given their existence, it is a different question, however, whether these sources of implicit reinforcement are sufficiently robust and consistent for a child to exploit them in syntactic development, or whether they are too sporadic and noisy to be useful (the latter view is emphatically endorsed, e.g., by Marcus, 1993). Because natural discourse is difficult to manipulate and control for experimental purposes, computational models of language acquisition provide an indispensable platform to shed more light on this issue.

In the simulations described so far, crucial aspects of child-parent communication have been left out of the picture. Specifically, the model did not implement the *goal-directed*, *interactive* and *dynamic* nature of human language learning. The model received linguistic input from its environment and was periodically tested in production on its progress in learning the target syntax. Along with the input it received meaning representations which encoded the semantic content of the overheard utterances. The model operated in a situated comprehension mode in which it had access to the meaning of its linguistic input. These learning conditions were motivated by the idea that children occupy a joint attentional frame with their parents and can partially infer the meaning of their parents' utterances from observed events in a shared visual scene. The testing procedure for the model, on the other hand, lacked important characteristics of human communication and learning. The utterances the model generated from meaning input were not addressing a discourse participant, they were produced in isolation. Thus, the model's productions served no communicative intention or pragmatic function and there was no incentive for the model to strive for communicative success. Although the model received input from its environment it did not engage in interactive conversation during acquisition. It's own productions were merely elicited for testing the state of syntactic development, but they were not directed at a listener from which it could receive the kind of differential feedback and positive reinforcement typical for child-parent interaction. Most importantly, the model's learning environment did not dynamically change in response to its own grammatical utterances. Instead, the model went through the same static training cycle and set of fixed input items again after testing, regardless of its current state of knowledge, and regardless of its own correct productions. The invariant input that the model repeatedly received did not reflect the changes and fluctuations in the distribution of child-directed speech that might ensue from successful child-parent communicative interaction.

In the condition of Section 7.3.2, the recursive Dual-path model's learning environment consisted of single-clause declaratives, single-clause polar questions, declaratives with relative clauses, and *wh*-questions with relative clauses. It was argued in the error analysis that the model was biased towards grammatical and against ungrammatical CPQCE and that its error profile matched that of children. The model spontaneously produced 8% fully correct CPQCE. However, it did not acquire CPQCE (or CPQRB, for that matter) to a degree of accuracy which was characteristic for the structures in the actual input. A reason might be that the model did not actively start to use CPQCE questions in communication and it did not receive positive reinforcement as a consequence of producing grammatical polar questions in the test phase. The environment did not reward the model in response to its correct productions by increasing the frequency of this structure in subsequent discourse. The model's bias towards structure-dependence was therefore not consolidated and converted into robust knowledge of CPQCE syntax. One approach to remedy this situation, which I explored here, was to feed back the correct CPQCE and CPQRB productions that the model accomplished during testing into the next training cycle. This approach is depicted schematically in Figure 7.6. Initially, the model received the same input structures as in the '*wh*-questions' condition. At some point in development the model produced grammatical CPQCE and CPQRB without prior exposure to these constructions. From this point onwards, it received its own correct productions in the subsequent training phase alongside the regular input. It was conjectured that interleaving the input with this feedback would help the model to acquire CPQ. Feedback was sparse at first, because the model only produced a few correct CPQCE/CPQRB structures in the beginning, but gradually increased over time as more correct productions accumulated in the feedback set.[17] The model had to test on

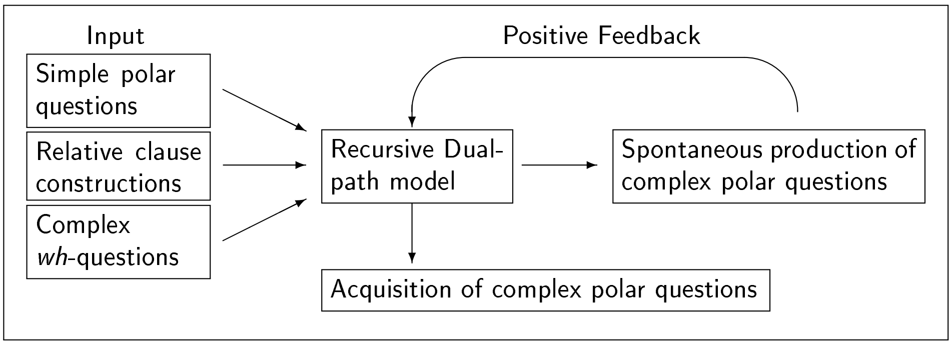

Input
Positive Feedback
Simple polar questions
Relative clause constructions
Complex *wh*-questions
Recursive Dual-path model
Spontaneous production of complex polar questions
Acquisition of complex polar questions

Figure 7.6: Positive reinforcement for complex polar question learning in the recursive Dual-path model.

[17] Duplicates of correct questions from different test cycles were filtered out to keep their overall frequency in the training environment low.

CPQCE/CPQRB structures with perfect accuracy before these could enter into the next training cycle. The model learned auxiliary fronting from the input structures it received and was then reinforced consolidating this piece of syntactic knowledge through exposure to its own correct productions. In this way we can model the idea that a human learner (a) may re-use constructions more frequently during development which she has used successfully in communication before, (b) receives positive reinforcement from discourse participants in response to grammatical and pragmatically adequate use, and (c) navigates through an adaptive and dynamically changing environment that reuses structures more frequently which have already been produced successfully by the learner before.

When I implemented this procedure, the model developed more robust representations of the syntax of auxiliary fronting than in previous conditions. Figure 7.7 shows the model's learning curves for all tested constructions. Single-clause utterances and SPQ were learned the fastest, followed by relative clause constructions and complex *wh*-questions. All input structures reached >95% sentence accuracy at the end of training. After exposure to approximately 10.000 sentences, the model began producing correct CPQRB structures and after 25.000 sentences it started to test positively on the novel CPQCE constructions. CPQCE developed slower than CPQRB but both structures reached >50% after training was complete.

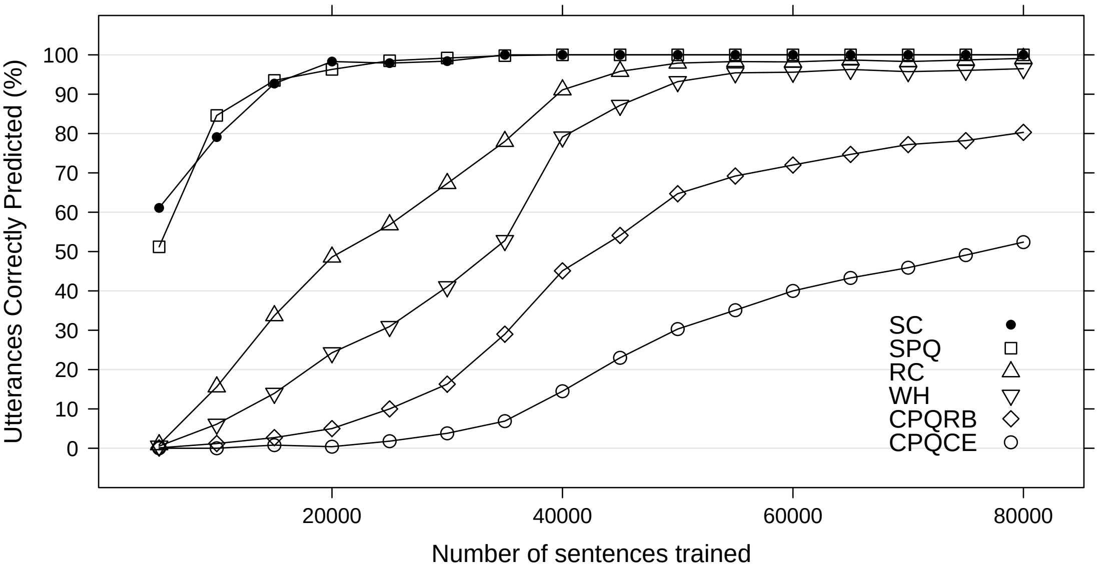


Figure 7.7: The recursive Dual-path model bootstraps into complex polar interrogatives; SC = simple-clause declaratives, SPQ = simple polar questions, RC = relative clause declaratives, WH = complex *wh*-questions, CPQRB = right-branching polar questions, CPQCE = center-embedded polar questions.

Not all model subjects bootstrapped equally well into the syntax of the CPQCE (mean accuracy: 53%, range: 20%– 87%, sd: 25%). There are several reasons for this.

To reach a high level of accuracy it was crucial that the model spontaneously produced correct CPQCE early in training to initiate bootstrapping. But not all model subjects started producing correct CPQCE at the same time. When bootstrapping was delayed in this way, the model was not able to fully recover later in training because the network's plasticity decreased over time.[18] Secondly, positive reinforcement was not instantaneous but occurred at some randomized point in time during the subsequent training cycle. It proved critical *when* reinforcement was given to the model. Even if it correctly produced CPQCE and was later rewarded for this behavior, it was not guaranteed to acquire this construction because at the time of reinforcement the model's early bias towards grammatical CPQCE could have been erased again by other linguistic input. And third, all structures which the model learned were represented over the same set of connection weights. CPQCE therefore competed with the similar CPQRB structures and the latter were easier to learn for the model because they were more similar to SPQ than CPQCE. Thus, CPQRB reinforcement started earlier in training than CPQCE reinforcement and this asynchrony could wipe out the model's bias to produce correct CPQCE early in training. Hence, CPQCE bootstrapping was delayed due to interference with CPQRB learning.

For completeness I report the production error for all four conditions I examined in one graph (Figure 7.8). Compared with the '*wh*-questions' condition, the error decreased significantly in the 'bootstrapping' condition. In particular, the production error for CPQCE dropped below 10%. When exposed to simple yes/no-questions, relative clause constructions, and complex *wh*-questions, the model was biased towards structure-dependent auxiliary fronting. When this bias was amplified by positive reinforcement for the model's spontaneous correct productions, it developed more robust syntactic representations which allowed it to bootstrap into the syntax of polar interrogatives. I will now take a closer look at these representations.

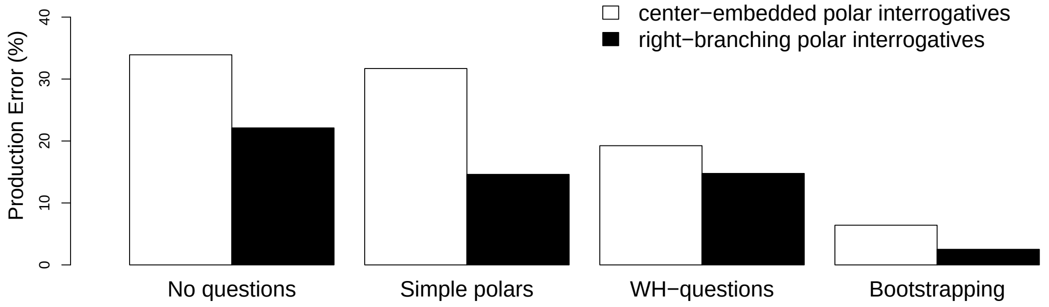


Figure 7.8: Mean production error for CPQCE and CPQRB in all four experimental conditions.

[18] The learning rate was set to gradually decay over training.

7.4 Structure-dependence revisited

In Figure 7.8, the model's performance was measured in terms of the production error for the entire test utterances. This procedure might yield potentially misleading information about differential performance across conditions, because a global decrease in production error may not adequately reflect learning of structure-dependent question formation. It is conceivable that the stepwise reduction of Figure 7.8 resulted from the model learning to appropriately sequence chunks of words within the tested utterances which are not relevant for the question whether structure-dependent auxiliary fronting was acquired. For instance, it may have resulted from improved learning of relative clause formation or post-relative clause sentence completion. The graph in Figure 7.8 does not reveal this kind of information and hence does not decide the issue whether the model respected the integrity of embeddings with regard to auxiliary extraction. I therefore examined only the relevant initial segments of CPQCE, up to and including the embedded clause auxiliary (e.g., *Is the boy that was…*). For these segments, essentially the same pattern of error reduction across conditions can be observed as for the complete utterances (Figure 7.9); a small difference between 'no questions' and 'simple polars', a larger reduction for '*wh*-questions', and the largest drop in the 'bootstrapping' condition. The model clearly improved on producing correct initial segments of CPQCE

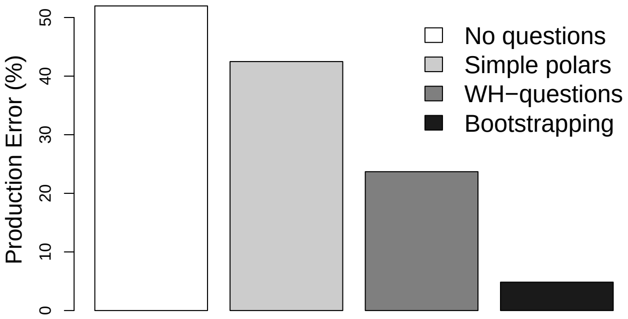


Figure 7.9: Production error for the initial segments of CPQCE.

across conditions. Again, however, this does not conclusively exclude the possibility that the model extracted the embedded auxiliary and merely improved on sequencing lexical items between the two auxiliary positions. Fortunately, the design of the recursive Dual-path model allows us to measure accuracy in terms of perfect match between a target structure and the actual output word-by-word. In this way we can determine the production rates of correct initial segments (*Is the boy that was…*) and incorrect initial segments (*Is the boy that kick…*) and this allows us to decide in which input condition the model favored which syntactic principle. At the end of training the model was given 500 test messages for CPQCE structures as input and this message was neutral between main and embedded clause auxiliary extraction. I measured how many of the model's utterances contained the intended initial segment AUX NP THAT AUX and how

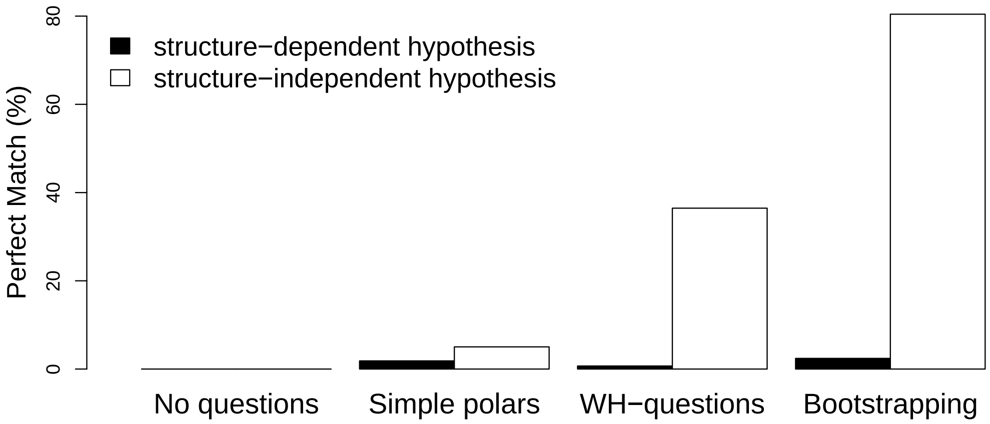


Figure 7.10: The recursive Dual-path model uniformly disregards linear auxiliary displacement and improves on hierarchical auxiliary fronting across conditions.

many contained the erroneous initial segment AUX NP THAT VERB in which the embedded auxiliary was extracted. The results for all four conditions are shown in Figure 7.10. The amount of correct initial segments increases from 0% in the 'no questions' condition to over 80% in the 'bootstrapping' condition and reaches 36% already in the '*wh*-questions' condition (without positive reinforcement). The number of incorrect alternatives in which the wrong auxiliary was fronted, on the other hand, remained constant and was minute across conditions. This indicates that the series of input conditions was successively more conducive to learning correct auxiliary fronting. It also suggests that *incorrect* productions did not reflect the model's preference for linear extraction in any of the conditions. In other words, in no condition did the recursive Dual-path model favor structure-independent auxiliary fronting. This is particularly interesting for the 'simple polars' condition in which the model was exposed to single-clause yes/no-questions, declaratives and relative clause sentences. Proponents of a universal grammar approach to structure-dependence often argue that a linear formation rule should be favored by a learner because it is consistent with simple yes/no-questions in the linguistic environment and at the same time simpler than a structure-dependent rule, which respects clause boundaries and the hierarchical organization of complex questions. Figure 7.10 indicates, however, that a data-driven learner may not adopt such a generalization from SPQ to CPQ even in the absence of any embedded questions in the input. This suggests that the dichotomy of structure-dependent versus structure-independent principles might not be the appropriate alternative to conceptualize the learning problem that children face for CPQ.

SRN and the recursive Dual-path model are sensitive to local substructure regularities in their sentence input. A possible explanation for the model's structure-dependent preference could therefore lie in the relative frequencies of substructures. The distribution of relevant substructures might support the structure-dependent hypothesis. Recall that the artificial language used to train the model had the verb stem and inflectional morphemes separated into two lexical items. In all conditions the model was trained on declarative relative clause constructions including center-embedded subject-relativized structures such as `the dog that chase -ed the cat is jump -ing`. Thus, the input included many sentences which contained the substructure NP THAT VERB. This is the critical substructure of a CPQCE initial segment that we expected the model to produce if it had adopted a structure-independent rule of question formation. Nonetheless, the model did not produce such sequences (Figure 7.10), despite being familiar with this it from exposure to other sentences which shared this substructure. Consequently, the model's behavior cannot be explained on the basis that it might have seen no example of the NP THAT VERB substructure but many examples of the correct NP THAT AUX substructure.[19] In the 'simple polars' condition, the input contained 1144 instances of NP THAT AUX VERB which supports structure-dependence and 930 instances of NP THAT VERB INFLECTION which supports structure-independence (averaged over ten model subjects). Thus the substructure distribution was slightly biased in favor of correct auxiliary fronting. When looking only at THAT AUX versus THAT no-AUX substructures, however, the model encountered 1369 instances of THAT AUX but 3237 instances of THAT no-AUX, where the pronoun was immediately followed by any lexical item other than an auxiliary.[20] From this point of view, the model's linguistic experience was strongly biased towards substructures which had no auxiliary following the pronoun and these substructures support linear auxiliary extraction. Similar distributions were found in the '*wh*-questions' and 'bootstrapping' conditions.

Kam et al. (2005) criticized other statistical approaches to auxiliary fronting (Lewis and Elman, 2001; Reali and Christiansen, 2005) by pointing out that high frequency of the bigram THAT AUX in the input might bias SRN to distinguish ungrammatical from grammatical CPQ. They argued that an SRN might simply exploit these bigram frequencies, rather than respect the hierarchical organization of complex questions, in order to learn auxiliary fronting. In addition, they argued that realistic English input would not support correct auxiliary fronting in terms of such bigram frequencies. In the input to the recursive Dual-path model, the bigram THAT AUX was not the most frequent bigram with an initial pronoun; for example the THAT ARTICLE was almost twice as frequent. Thus, if sensitivity to these bigrams was the prime determinant of the model's behavior, we would not expect a preference for structure-dependence. When producing an utterance from the novel CPQCE message input, the model had to either produce an auxiliary or a verb stem after the pronoun. But both bigrams THAT AUX and THAT VERB were almost equifrequent and in none of the four conditions did the model produce the incorrect substructure in more than 2.4% of the tested utterances. Given the overall parity of 'substructure evidence' which the model received, this suggests that

[19] Separating verb stem and inflectional morphemes pays off here, because otherwise the embedded verb forms in subject-relativized declaratives and incorrect polar interrogatives would systematically differ.

[20] THAT AUX substructures occurred in subject-relatives with progressive aspect and all subject-relativized passives, THAT NO-AUX substructures were found in all object-relativized clauses and in subject-relatives with simple aspect.

the structure-independent principle of auxiliary fronting might just not appear simpler and more preferable to a domain-general learner (such as a backpropagation network) than the structure-dependent principle.

Unlike SRN, the recursive Dual-path model does not learn from word sequences alone, but also receives a meaning representation for word sequences as input. The model's behavior is not fully determined by substructure frequencies in the training corpus. Messages used generic semantic features to signal the number and relative prominence of participants in the event semantics (see Chapters 3 and 4). They are not idiosyncratic but composed from a few features in a systematic way. This encoding creates semantic similarities and partial meaning overlap between different constructions in the training language. These semantic similarities help the model to produce novel utterances in generalization tasks such a polar interrogative learning. In training, the

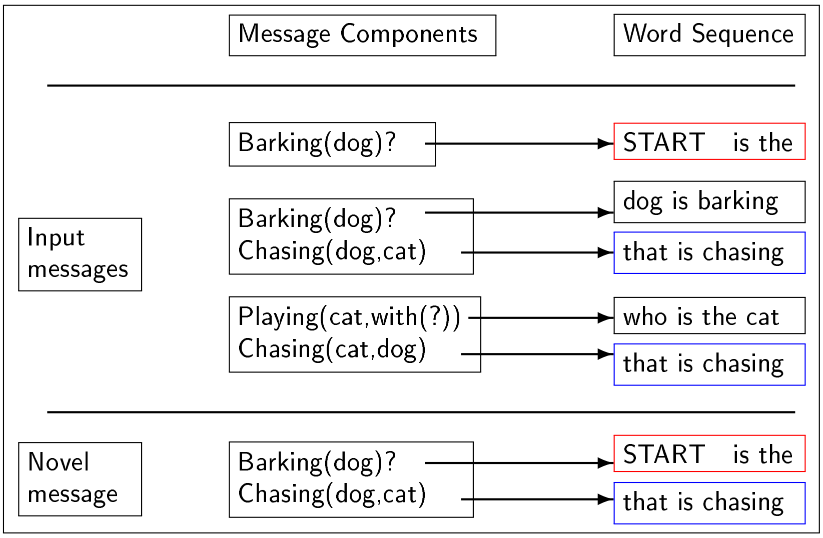


Figure 7.11: Different components of the meaning representations control different subsequences of words in the target structure.

model learns to associate subparts of a sentence with the proposition in the message that controls it. For example, from simple-clause messages the model learns to sequence participants in atomic events such as dative transfer (*dog gives toy to cat*). Other features of the message control the position of relative clauses and the semantic role of the head noun in the relative clause. By attending to those features, the model learns how the generation of embeddings is controlled in the message. When presented with a message for a novel structure, the model can use substructure regularities in its input message and combine these regularities to generate sentences with a novel hierarchical organization (e.g., additional embeddings). In learning the syntax of auxiliary fronting, regularities in the message play a crucial role. Figure 7.11 illustrates the process of recruiting semantic information from familiar message-sentence pairs in the construction of novel utterances. Sentence meaning helps the model to segment utterances into the parts that correspond to the main clause (e.g., *Is the dog barking?*) and the part related to the embedded clause (e.g., The dog *that is chasing the cat* is barking). Thus, the auxiliary in the main clause is controlled by a different part of the message than the auxiliary in the embedded clause. In training on SPQ and complex *wh*-questions, the model learns to associate the clause-neutral question feature with the main clause part of the message. When a new complex question is created, the auxiliary that is controlled by the main clause message is shifted to the front. In this way the system learns that picking the closest auxiliary is not appropriate.

7.4.1 Hidden layer analysis

The relative frequencies of bigrams in the input did not explain the model's performance on novel CPQCE. Since these relative clause internal bigrams were the critical subsequence which distinguished grammatical from ungrammatical CPQCE initial segments, this indicates that the model might not represent knowledge of auxiliary fronting as an operation on linear sequences of lexical items. Rather, it seems that the model is representing the hierarchical structure of complex sentences and respects the clausal integrity of the embedding when extracting the auxiliary. It was argued above that the model learned which features of the event semantics controlled which aspects of the hierarchical organization of complex sentences from message-sentence pairs in training. Because novel messages were sufficiently similar to experience, the model could built novel structures from relevant components of familiar messages. Thus, generalization to CPQ was enabled by similarity-based meaning-to-form transduction.

Hierarchical structure

If this is the correct explanation, we should be able to observe reflections of semantic similarity and dissimilarity in the model's internal representations. To test this, I analyzed the HIDDEN-layer activation states of the model during the production of CPQCE and CPQRB structures, respectively. Specifically, I looked at the two sentences

(24) a. `Is the father that was push -ing a brother walk -ing ?` CPQCE
b. `Is the father push -ing a brother that was walk -ing ?` CPQRB

Both sentences share the lexical items `is the father push -ing a brother walk -ing` but the common subsequence `that was` is displaced. The subsequence `is the father` belongs to the main clause of both sentences. In sentence (24-a) the subsequence `push -ing a brother`, however, is part of the center-embedded relative clause, in sentence (24-b) it is part of the main clause. In both cases, `the father` is the agent of a transitive action and `a brother` is the patient. Similarly, the subsequence `walk -ing` is the verb of the right-branching relative clause in (24-b) whereas in (24-a) it is the main clause verb; the two occurrences of `walk -ing` take different subjects. If the model represents the syntax of auxiliary fronting as a linear operation over strings of words we would expect that it represents both sentence types in a similar way, at least with respect to their common subsequences. In other words, we would expect that the model does not distinguish the distinct hierarchical organization of these sentences in terms of clausal units and dependencies. If, on the other hand, the model learned auxiliary fronting from similarities in semantic representations it should represent the difference in clausal structure between the two sentences. Then we should be able to detect this difference by visualizing the internal representations the model has developed at the end of training.

For a model subject trained in the '*wh*-questions' condition, I recorded HIDDEN-layer activation for the entire set of test sentences. As each word of each test sentence was passed through the model a snapshot was taken of the HIDDEN-layer state, yielding roughly 15.000 60-dimensional vectors. A principal components analysis was performed to identify the HIDDEN-layer dimensions which explained the bulk of variation in this data. I also recorded activation states for questions (24-a) and (24-b). These particular questions were selected because the model produced both with perfect accuracy. The activation states obtained from the test questions were then plotted in terms of two principal components. Figure 7.12 depicts the HIDDEN-layer states for both questions with respect to principal components 2 and 4. One can think of these graphs as a sentence's trajectory through HIDDEN-layer space, visualized through data compression and projection onto a new coordinate system. In Figure 7.12, both trajectories start out in the same region of state space while the model produces the common main clause sequence `Is the father`. At the verb `push -ing`, however, a bifurcation of trajectories occurs as this verb belongs to different clauses within the two structurally distinct questions. The separation of trajectories further amplifies for `a brother` which is the embedded clause direct object in (24-a) and the main clause direct object in (24-b). As the sentences are completed the trajectories remain separated but approach each other again towards the end of sentence marker (labelled 'eos').

This analysis supports two conclusions. First, the recursive Dual-path model represents distinct types of relative clause constructions very differently in HIDDEN-state space although there is maximal lexical overlap. Secondly, the model represents phrasal units with the same constituents which occupy the same syntactic and semantic roles in different regions of state space depending on whether they occur in the main clause or an embedding (e.g., `a brother`). This suggests that the model is not representing knowledge of auxiliary fronting as linear operations on strings of words but in terms of the hierarchical organization of complex sentences into clausal units.

In a recent paper, Perfors, Tenenbaum, and Regier (2008) argued that connectionist models of syntactic development are not making a meaningful contribution to the auxiliary fronting debate due to the nature of their internal representations. Since these models are not representing hierarchical sentence structure, as they claim, they cannot on principle help to investigate the question whether the structure-dependent syntax of polar interrogatives is learned through domain-general mechanisms, or innate and language-specific. While this might be a valid point for SRN which learn from word sequences alone and depend on substructure frequencies in order to acquire auxiliary

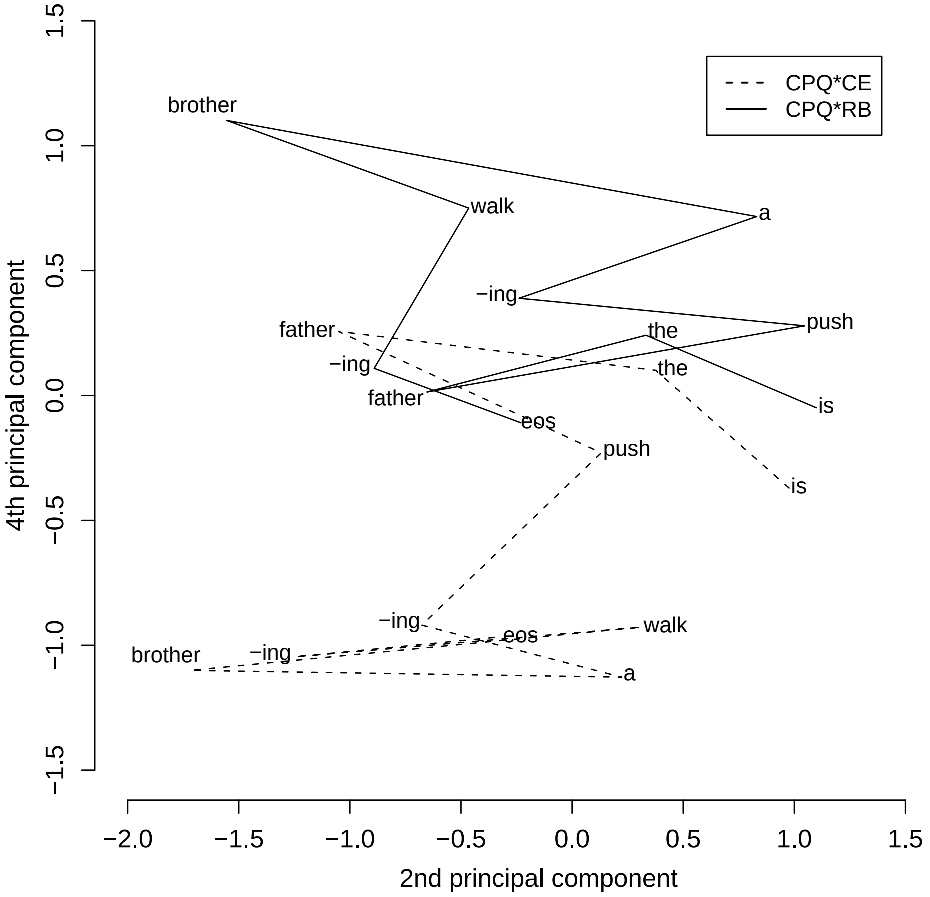


Figure 7.12: Principal components analysis of the model's internal representations for the common subsequence `is the father push -ing a brother walk -ing` of questions (24-a) and (24-b).

fronting, the recursive Dual-path in addition draws on semantic input. The model learns to associate subparts of this input with distinct clauses and this allows the model to organize complex sentences into clausal units with distinct HIDDEN-state space representations of main and embedded clause. Subsequences common to different types of CPQ are represented in different regions of state space.[21] As a consequence, auxiliary fronting in the Dual-path model is truly structure-dependent and not rooted in frequent substructures. These findings cast some doubt on the validity of the judgement voiced by Perfors et al. (2008).

[21] Note that the logic of principal components analysis only requires the existence of principal components which distinguish constructions for this argument to be sound, in our case the 2nd and 4th component.

Generalization by analogy

It was a main result that the recursive Dual-path model was able to spontaneously produce correct CPQ, displaying an error profile that matched experimental data from children, when exposed to simple questions, relative clause constructions, and complex *wh*-questions. Although these questions differ from polar questions in their syntax and surface form, they aided the model in developing representations which allowed the transfer of auxiliary fronting. This suggests that the model acquired a notion of structure-dependence which generalizes to novel constructions by analogy, with semantic similarities being the basis of this analogy. According to Tomasello (2003), analogy is one of the crucial mechanisms of child language acquisition—besides entrenchment through repetition and preemption of ungrammatical forms (see also Goldberg, 1999; Israel, 2002). On this view, analogy is a source of grammatical generalization (but also overextension) when there is a good structural mapping in terms of linguistic form and communicative function between whole utterances or constructions. If this is true, deep structural relations as well as surface similarities should be reflected in the linguistic representations of constructions acquired by analogy. We can test this idea again by visualizing the internal representations of the trained model when processing the complex questions

(25) a. `is the father that was push -ing a brother walk -ing ?`
b. `who is the father that was push -ing a brother walk -ing with ?`

If the model draws on the semantic similarities between the two types of questions to learn the syntax of polar interrogatives we should expect the corresponding HIDDEN-state space trajectories to be similar. I plotted the model's activation states for the sentences (25-a) and (25-b) with respect to the same two principal components as in the previous Section (Figure 7.13). Although the resulting trajectories do not match perfectly they are very similar qualitatively and could be made nearly congruent by shifting the lexical subsequence they have in common. The largest offset between trajectories occurs at the sentence initial segment `who is` and the sentence final segment `walk -ing with` which distinguish the two question types in terms of surface form, argument structure and communicative function. This suggests that knowledge of auxiliary fronting transfers from *wh*-questions to yes/no-question by analogy. Semantic similarities between these questions in the message drive this analogical process and lead to similar internal representations for the two constructions.[22]

[22] Note that this is an inference to the best explanation not a consequence of this analysis. Note also that it is not sufficient, in general, for this kind of argument to look at two principal components only. The representational similarity was investigated for the first fifteen principal components which explained 75% of the variance in the data and trajectories did not diverge along these dimensions.

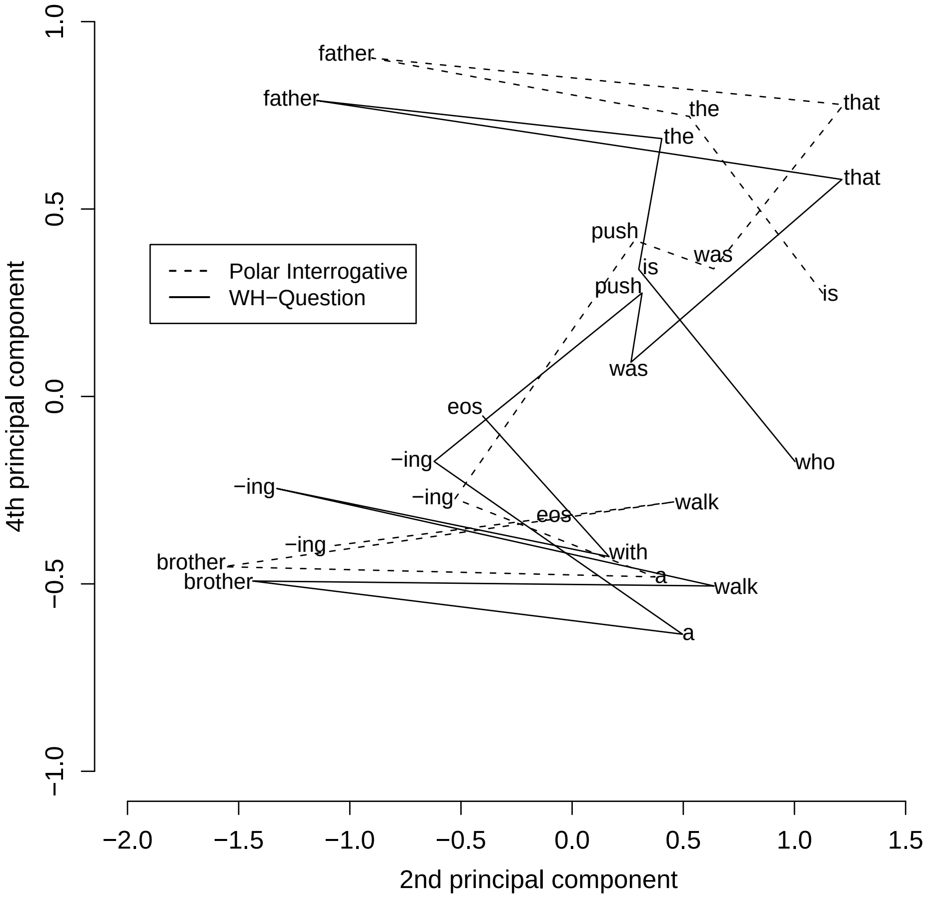


Figure 7.13: The internal representations of complex yes/no-questions and complex *wh*-questions display qualitative similarity in HIDDEN-layer space.

Questions and declaratives

Connectionist models of auxiliary fronting have been criticized for not learning the right kind of mapping between declarative and interrogative expressions (Frank et al., 2006). According to these authors, such models treat both types of sentences independently and the representations they develop do not reflect their close correspondence.

On standard views in syntactic theory, declarative sentences are transformed into questions by primitive operations such as movement, rearrangement or deletion of constituents. Tomasello (1992) argued that this perspective is not accurate developmentally because some children learn *wh*-questions before they learn word combinations that license transformation. Moreover, the majority of children's early *wh*-questions is highly formulaic, e.g., *what NP doing?*, *where is THING?*, etc. (Dabrowska, 2000). In construction grammar, questions are treated as separate constructions with a characteristic communicative function, which are not derivative but have to be learned just like any other construction. Nonetheless, in this framework question constructions are combined from a large number of other constructions, the subject-auxiliary inversion construction, the relativization construction, and so forth. For instance, in the polar question *Is the man giving the dog a toy?* it is the same ditransitive construction for declaratives that is used in a question form (Goldberg, 2006). The recursive Dual-path model reflects this approach in that the meaning representation for questions is built from declarative messages and question features. Because the semantics of questions is compositional in this manner, we should observe a close correspondence in the model's representations of these two constructions. Figure 7.14 illustrates this correspondence for the sentences

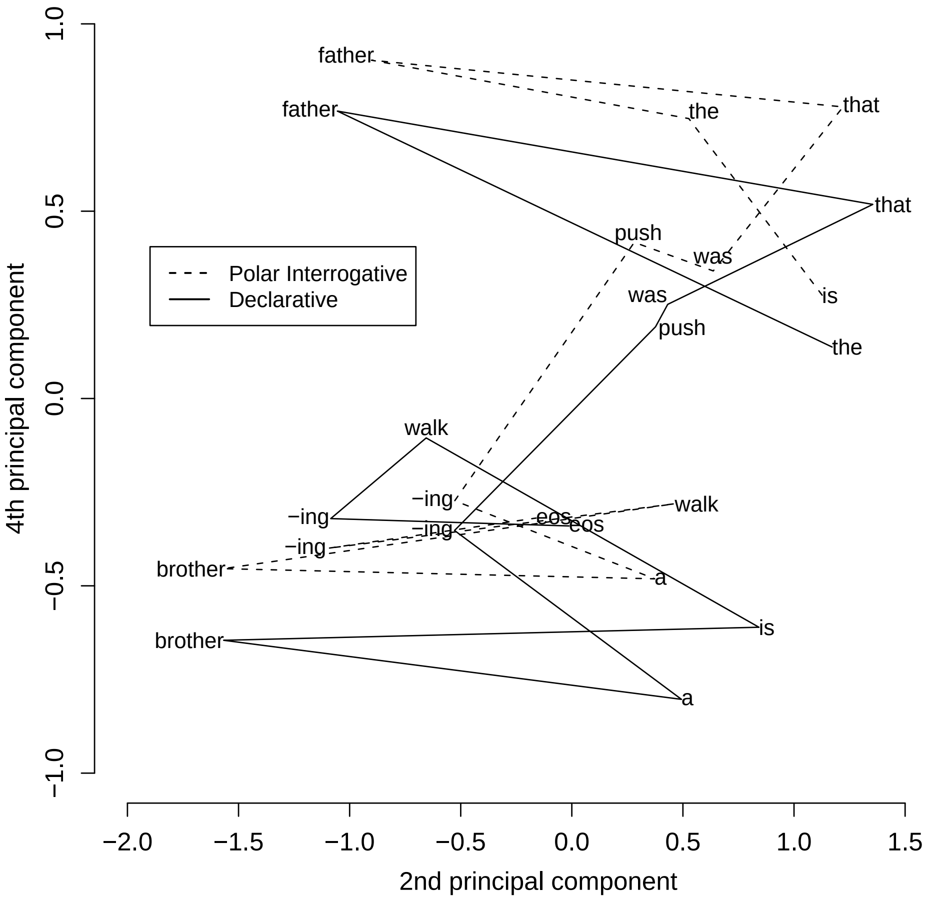


Figure 7.14: HIDDEN-layer representations of polar interrogatives and declaratives reflect their structural and semantic correspondence.

(26) a. is the father that was push -ing a brother walk -ing ?
b. the father that was push -ing a brother is walk -ing .

Again, the two trajectories match qualitatively with the largest divergence occurring at the subject NP determiner. This suggests that the model represents the structural and semantic relationship between questions and declaratives. The semantics of both constructions in the model's message differed only in one feature and this similarity in meaning caused the model to develop very similar representations of the two structures. The representational correspondence is induced by semantic similarity, not by a syntactic operation which transforms one structure into another.

7.5 Discussion

The aim of this chapter was to shed some light on the data-driven learnability of auxiliary fronting in polar interrogatives. After discussing the logic of poverty of the stimulus arguments and the hidden assumptions behind such arguments, I formulated three empirical hypotheses how polar interrogatives might be learned in the absence of positive evidence. These hypotheses have been tested in two similar computational learning mechanisms.

According to Hypothesis-3, distributional regularities in the input to a statistical learning mechanism are sufficient to acquire knowledge of auxiliary fronting in the absence of complex polar interrogatives in linguistic experience. I reviewed two studies, Lewis and Elman (2001) and Reali and Christiansen (2005), which explored this hypothesis in the framework of simple recurrent networks. These studies suggest that in principle SRN are able to avoid structure-independent generalizations based on distributional properties of the input. I argued that the Lewis and Elman (2001) model learned from input which was deliberately impoverished to bring about this behavior whereas, on the contrary, the Reali and Christiansen (2005) model learned from input which was unnaturally enriched in that the specific corpus tagging provided the model with substructure information which might not be available to a human learner. I also argued that the range of tested structures was too narrow in both studies to warrant general conclusions about the data-driven learnability of complex polar questions. Nonetheless, both studies are important in that they indicate that a general-purpose learner might become biased against structure-independent errors without corrective evidence against overgeneralization in the form of complex polar questions in the input.

In my own modelling approach I worked with the recursive Dual-path model which is one of the few explicit models of language production that uses meaning for syntactic development. This model learns associations between parts of semantic representations and subsequences of words, and it can combine these regularities in novel ways. In the previous chapter, for instance, it was demonstrated that this mechanism could explain the generalization of familiar words to novel slots, and the generalization of subsequences to novel embeddings. I examined two hypotheses about auxiliary fronting in this model which both involve the generalization of regularities from simpler and similar structures to novel polar interrogatives. Hypothesis-1 suggested that these structures might be learnable from exposure to single-clause declaratives, simple polar questions and relative clause declaratives in a constructive, bottom-up manner. The idea was that subject-auxiliary inversion is learned from simple questions and relativization from complex declaratives and that this knowledge could be combined by the model to form complex questions, because it learned how subparts of complex sentences were controlled by subparts of the semantic input. Since all input is processed over one set of connection weights, similarities between experienced and novel messages would allow the model to correctly produce novel complex questions. This constructive approach is similar to a proposal of Ambridge et al. (2008) that children might be able to learn the syntax of complex questions by learning to substitute complex NPs (*the dog that is chasing the cat*) for simple NPs (*the dog*) in simple questions (*Is the dog running?*), based on surface distributional properties of the input and the functional-semantic similarity of these types of phrases. Within the recursive Dual-path model framework I did not find evidence which supports this proposal; simple polar questions and relative clause constructions in the input were not sufficient for the production of complex polar questions or even a substantial amount of correct initial segments. Negative results from a computational model can of course never invalidate a particular theory of acquisition. Rather, they indicate that the learning mechanism may not be adequate to implement this theory or that the modeller did not identify the optimal learning conditions for the system.

Hypothesis-2 conjectured that polar interrogatives might be learnable if in addition to the input under Hypothesis-1 the model was exposed to complex *wh*-questions. In this condition, I found that the model was able to generate correct complex polar questions without having seen this structure in training. Moreover, the model's error profile matched the developmental data of Crain and Nakayama (1987) and Ambridge et al. (2008) in that the most frequent error type involved the duplication of the main clause auxiliary after the embedding. This indicates that the model required exposure to other complex questions in which subject-auxiliary inversion was non-local, i.e., occurred between non-adjacent constituents, in order to learn auxiliary fronting in polar questions. The model's behavior across these two condition suggested that complex *wh*-questions might have been necessary in the input to block structure-independent generalization from simple to complex polar questions. If this was the correct explanation, we should have observed many initial segments with displaced embedded auxiliaries in the condition of Hypothesis-1. However, this was not the case. In neither condition did the model produce such ungrammatical initial segments which indicates that complex *wh*-questions were not required to inhibit overgeneralization. The effect of simple polar questions in the input, apart from providing examples of subject-auxiliary inversion, was visible in that the model produced less errors in right-branching than in center-embedded polar questions in both conditions, since simple polars were more similar to main clauses of the former than the latter structure.

Neural network models—such as SRN or the recursive Dual-path model—are better at learning local regularities than long-distance dependencies. In simple polar questions, auxiliaries and their corresponding verb form are separated only by the subject NP. In complex polar questions these constituents are separated by more material, the main clause subject NP plus an embedded clause. Such a long-distance separation is never witnessed by the model in the condition of Hypothesis-1 and the model forms strong statistical expectations that auxiliaries in questions are always followed by the subject NP and then the verb form. In the condition of Hypothesis-2, on the other hand, complex *wh*-questions provide positive evidence that in questions additional material can intervene between the subject NP and the dependent verb form and this weakens the transitional probability between these constituents. When presented with a message for a novel complex polar question, relativization of the main clause subject NP inside a question is a grammatical option for the model under Hypothesis-2, whereas it is ruled out under Hypothesis-1 by the model's linguistic experience. In other words, complex *wh*-questions appear to be necessary in the input to override the statistical expectations of the model's sequencing system. The model needs to witness some input structures in which the main clause auxiliary and the verb form are separated by more material than just the subject NP in order to be able to combine the subject-auxiliary construction with the relativization construction in generating novel complex polar questions.

The recursive Dual-path model approach to auxiliary fronting can be criticized on the basis of the input the model receives in learning. It might be argued, for instance, that meaning input which corresponds to clauses in the target structure preempts the problem of learning structure-dependence because the hierarchical organization of complex questions is already encoded in the conceptual structure of the message. It should be noted, however, that the model needs to *learn* which components of the message control which substructure of an utterance. For example, it needs to learn which parts of the message control the generation of embeddings and which parts control the sequencing of participants inside an embedding. Once the model has learned to interpret its semantic input in this way, it can produce relative clause declaratives and my approach explicitly assumed that a notion of relativization in declaratives precedes learning the syntax of complex polar questions. If it was not assumed that children in the Crain and Nakayama (1987) study were able to comprehend relative clauses, this would entail that the very instructions of the experimenters were unintelligible to their subjects; the study would be methodologically flawed and pointless. Secondly, the model was never exposed to messages for complex polar questions in training. These messages shared semantic features with the corresponding declarative, e.g., features which controlled the embedding. Nonetheless, these messages were novel in that they combined declarative features with question features and this combination was not experienced by the model in learning. Based on experience with other message-sentence pairs, the model had to utilize semantic components in the novel messages in the appropriate way to generate polar questions. It was precisely the point of my approach that positive evidence from which the syntax of polar questions could be learned was not restricted to simple and complex polar questions alone but included complex declaratives and other types of questions as well. And third, recall that polar question features in the message were entirely neutral with respect to clause type (main or subordinate). When the model received message input for a complex polar interrogative, the question feature could refer to both clauses. The model was not biased by the message input to select a particular clause as the locus of subject-auxiliary inversion and Figure 7.10 (page 238) demonstrates that nevertheless it did not select the embedded clause auxiliary in *any* of the learning conditions. For these reasons, the conceptual structure of the model's semantic representations does not beg the question of learning structure-dependent auxiliary fronting.

A second point of criticism might be raised against the particular input distribution in the condition of Hypothesis-2 in which 10% of the model's training items were complex *wh*-questions. This amount of exposure by far exceeds the most optimistic frequency estimates for such questions in child-directed speech. I critically evaluated the studies of Lewis and Elman (2001) and Reali and Christiansen (2005) based on the kind of input they were exposed to but it seems that the recursive Dual-path model approach does not fare any better in this respect. In my approach, the model received training with an artificial, English-like language similar to the Lewis and Elman (2001) study. Unlike their input environment, however, my language was combinatorially complete in that every combination of basic constructions was admissible. Every syntactic role could be modified and relativized, both in declaratives and *wh*-questions, and every combination of tense and aspect could occur in the two clauses of a complex sentence. Critically, no grammatical structure that this language could generate was omitted from training for the purpose of creating conducive substructure frequencies. Compared with the Reali and Christiansen (2005) study, my language lacked the structural variation and noisiness of their training corpus. On the other hand, input tagging was not required and this ensured that the model's performance was not dependent on substructures which may have resulted from a specific tag set in their study. Thus, although the input to the recursive Dual-path model was artificial and lacked a realistic distribution, I believe my results rest on less controversial assumptions about the learning environment.

It was shown that the recursive Dual-path model was able to spontaneously produce grammatical complex polar questions without exposure to these structures in the input. At the same time, the model did not produce a substantial amount of ungrammatical, structure-independent initial segments in any input condition. This suggests that the *absence* of input patterns in which auxiliaries are extracted from an embedding provides a strong cue for the model that structure-independent generalizations are ungrammatical. Since the displacement of auxiliaries across clause boundaries is never witnessed by the model, it does not entertain this possibility when generating complex polar questions.[23] It seems that standard accounts of the learning problem for auxiliary fronting do not factor such 'evidence from absence' into the equation. The behavior of my model indicates that such negative evidence by itself might be sufficient to block overgeneralization. Nativist arguments for the innateness of structure-dependence might therefore be fundamentally flawed. These arguments construe the learning problem as a binary choice between right and wrong generalization of which only the latter is licensed by the input. Consequently, children *should* overgeneralize and since they apparently do not, the innateness of structure-dependence appears to be the only viable explanation. But if there is no reason to believe that children should overgeneralize to structure-inde-

[23] A similar point is made in Regier and Gahl (2004) with respect to their Bayesian learning model for anaphoric *one*.

pendent auxiliary displacement based on their linguistic experience, a crucial premiss of the argument breaks down and the poverty of the stimulus quandary becomes a straw man which is not in need of explanation. Instead we should attempt to re-conceptualize the learning problem in ways which suggest strategies how structure-dependence might be learnable and investigate these strategies in computational mechanisms. The recursive Dual-path model approach to auxiliary fronting conjectured that the syntax of polar questions is learnable in a constructive fashion from simpler and similar building blocks such as simple questions, relative clauses and other types of questions whose occurrence is warranted in child-directed speech. A crucial assumption of this approach, which is shared by many psycholinguists (e.g., Pinker, 1984, Tomasello, 2003), was that meaning plays an important role in syntactic development. Language acquisition, on this view, is not reducible to syntax acquisition but is driven by the objective to 'make sense' in learning to successfully map between meaning and form. Because the meaning of novel questions was sufficiently similar to the meanings of constructions which the model experienced in learning it was able to generate complex polar interrogatives from meaning representations without direct exposure to these structures. The results I presented from this approach suggest that the *structure of meaning* may obviate the need for innate syntax-specific knowledge in the acquisition of adult-like language abilities.

Chapter 8

The accessibility hierarchy

The accessibility hierarchy (AH) stratifies relative clause constructions in terms of the relativized NP's syntactic role (Keenan and Comrie, 1977) and this is considered to be an implicational universal in typology. I explore here an account where similarity and frequency of substructures in the input are the primary sources of the AH. This input-based account is consistent with usage-based syntax acquisition work (Diessel and Tomasello, 2005). The recursive Dual-path model was taught an English-like language through exposure to message-sentence pairs and its behavior during development displayed the AH ordering. I was able to manipulate and remove this ordering by varying properties of the input, and this suggests that patterns of interference and facilitation among structures can help to explain the AH in processing and development within a connectionist learning model.

8.1 Introduction

Language universals are important in theories of language, because they suggest that there are aspects of languages which may be innately endowed. Theoretical accounts of language universals sometimes argue that they arise from the nature of an innately specified language processor. Another possibility, that I examined here, is that these universals arise from the mechanisms of a language learning system. One important syntactic universal in linguistic typology is the accessibility hierarchy of relative clause constructions. English relative clauses can be distinguished based on the grammatical function of their head noun in the relative clause. For example, in the sentence `the boy that runs`, the constituent `boy` functions as the subject of the intransitive clause and I label this construction as an S-relative. Sentences with transitive subjects relativized are called A-relatives, and so forth (other types of relative clauses are presented in Table 8.1). I will use this labelling throughout to refer to sentence tokens in the example column. Keenan and Comrie (1977) sampled relative clause constructions from 50 languages and based on this data they formulated an implicational universal for all languages. If a language knows a construction to relativize subjects (S + A) and any other grammatical

Relativized role	Example	Label
Subject intransitive	... the boy that _ runs	S
Subject transitive	... the boy that _ chased the dog	A
Direct Object	... the cat that the dog chased _	P
Indirect Object	... the girl who the boy gave the apple to _	IO
Oblique Object	... the boy who the girl played with _	OBL
Genitive	... the man whose _ cat caught the mouse	Gen
Obj. of Comparison	... the cat that few are cuter than _	OComp

Table 8.1: Summary of English relative clause constructions.

role in the ordering

$$(S + A) > P > IO > OBL > Gen > OComp$$

then it can relativize any position in between using the same construction. Any reordering of relative clause types would invalidate this implication. In linguistic typology this ordering is known as the noun phrase *accessibility hierarchy* (AH). Syntactic universals such as the AH are considered prime candidates for being expressed in universal grammar.

A syntactic construction (or 'relativization strategy') to relativize subjects is also called a *primary strategy*. In English, most hierarchy positions can be relativized by using the relative pronoun `that` and omitting the head noun in the relative clause (see S-, A-, and P-relatives in Table 8.1). Alternatively, English can use other pronouns (`who`, `whose`, `whom`) for the AH positions (except OComp). Both constructions are primary strategies. Other languages, however, require more explicit relativization strategies on lower positions of the hierarchy. For example, OBL-relatives in Welsh need to retain an anaphoric pronoun (`it` in the example below) in the canonical position of the head noun in the relative clause:

(1) y llfyr y darllenais y stori ynddo
the book COMP read.1SG.PST the story in.it
the book in which I read the story

Pronoun retention is a *secondary strategy* in Welsh because it cannot be used for subject relativization. According to Comrie and Keenan (1979), for each position on the AH there is a language which knows a primary strategy for this position but requires a secondary strategy on any lower position.

Keenan and Hawkins (1987) speculated that this hierarchy may be rooted in processing difficulties. The lower a construction occurs on the AH, the more difficult it is to process. If in some language a relativization strategy works for position k of the hierarchy and for subjects (primary strategy), then we would expect that the language reuses this strategy on positions above k where relative clause constructions become successively easier to process. Conversely, if some relative clause construction is difficult to process, a language user may have to employ a different, more explicit relativization strategy on positions below k to facilitate production and comprehension (secondary strategy). In this way, both AH constraints on relativization could be explained in terms of cognitive economy. To test the idea that the AH is correlated with processing difficulty, Keenan and Hawkins conducted an experiment in which subjects had to first comprehend and then reproduce different relative clause types. Short-term memory effects were eliminated in that subjects had to repeat random sequences of digits before production. Repetition accuracy scores for the different construction types on the hierarchy are shown in Figure 8.1. Although not all contrasts were significant, they found

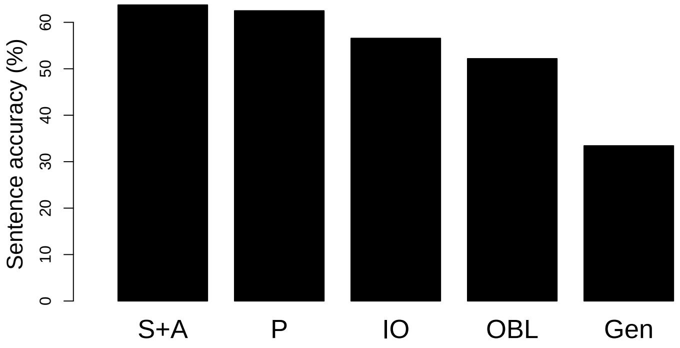


Figure 8.1: Adult relative clause comprehension accuracy matches the AH ordering (Keenan and Hawkins, 1987).

that the order of difficulty in adults qualitatively matched the accessibility hierarchy ordering (with the exception of the OComp construction not reproduced here).

A number of processing accounts have been proposed to explain this kind of data, based on the syntactic structure of relative clauses and/or working memory limitations. For instance, Hawkins (1994) defined a metric for the processing difficulty of relative clause types in terms of phrase-structure tree complexity. The more embedded the trace of the head noun is within the relative clause, the more structurally complex and hence more difficult it is to process. This complexity metric crucially depends on the syntactic analysis of relative clauses in some phrase-structure grammar, e.g., the principles and parameters framework.

According to Hale (2006), the AH in sentence processing can be explained as a function of entropy reduction in incomplete parse trees. The idea is that as a sentence is processed incrementally in comprehension, some words carry more information about the syntactic structure of the rest of the sentence than others. The parser has to project the structure of the sentence based on the words it has encountered already and the difficulty in this task is proportional to the amount of conditional uncertainty at each sentence position. Consequently, some relative clause constructions are harder to parse than others and the predictions derived from this metric correlate well with the data of Keenan and Hawkins (1987).

A prominent theory which stresses the role of working memory limitations in relative clause processing is Just and Carpenter (1992). They devised an activation-based model of reading times for English relative clause constructions which shows that cognitive capacity constrains comprehension. This model assumed that working memory is simultaneously involved in storing partial representations of incomplete parses and in integrating incoming constituents into these representations. The effect of constraining working memory resources in the model is that object-relativized structures are more difficult to process than subject-relativized structures.

One of the most influential proposals, the *dependency-locality theory* of Gibson (1998), suggests that the hierarchy can be accounted for by combining two factors, the varying distance between the head noun of the relative clause (called 'filler') and the canonical position of the head noun in the relative clause (called 'gap'), and the number of incomplete syntactic dependencies at each sentence position.[1] Both factors, it is claimed, are taxing the human sentence processor because the filler has to be kept in working memory until it can be integrated at the gap position, and similarly, open dependencies induce a memory cost until they are resolved. In case of sentences with just one relative clause, which are studied in this chapter, dependency-locality theory reduces to filler-gap distance as the crucial factor, the number of incomplete dependencies is negligible. Thus, Gibson's theory would adequately predict, for instance, that S-relatives are easier to process than P-relatives (see Figure 8.2), because the distance between the filler (in this case man), and the gap (indicated by the underscore in the examples) is larger in P-relatives than in S-relatives. The dependency-locality approach is also relevant for explaining why OBL- and IO-relatives are lower on the processing hierarchy than P-relatives.[2]

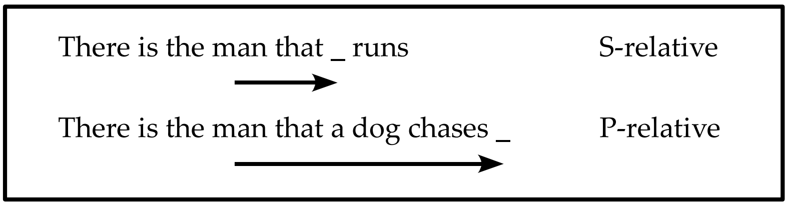

There is the man that _ runs
S-relative
There is the man that a dog chases _
P-relative

Figure 8.2: Filler-gap distances in S- and P-relatives.

[1] Filler-gap distance as a processing factor was first described in Wanner and Maratsos (1978), the second factor is a variation on the number of incomplete phrase structure rules held in working memory, mentioned in Chomsky and Miller (1963).

[2] In the psycholinguistic literature many more theories have been proposed which partially explain processing differences between constructions in the AH, for instance the conjoined clause hypothesis (Tavakolian, 1981), the non-interruption hypothesis (Slobin, 1973), the parallel function hypothesis (Sheldon, 1974), the NVN-schema hypothesis (Bever, 1970), and the perspective keeping hypothesis (MacWhinney and Pléh, 1988). For a summary and discussion see Diessel (2004) and Chapter 5.

8.2 The accessibility hierarchy in development

There are several aspects of AH behavior which are not addressed by filler-gap distance processing accounts. First, these accounts may not make the right cross-linguistic predictions for relative clause processing. German relative pronouns, for example, are marked for gender, case, and number. Hence, in most sentences with relative clauses the grammatical role of the gap is resolved at the pronoun position already and the filler need not be kept in working memory until it is being integrated at the gap site. Secondly, these processing accounts pertain to comprehension, but presumably in production no filler integration is required at the gap position because the speaker mentally represents the intended message in a definite, unambiguous way; she knows the syntactic role of the head inside the relative clause from sentence onset. A third issue which has not been examined carefully concerns the relationship between filler-gap distance processing accounts and language acquisition. It is not clear whether such accounts would work with the incomplete syntactic representations that children are using. These theories might not predict the hierarchy behavior in development. On the other hand, if children are not making adult-like predictions based on adult-like syntactic representations stored in working memory, then they might not exhibit adult-like AH behavior. In a sentence repetition study with English children [4;3-4;9], however, Diessel and Tomasello (2005) found that the order of relative clause acquisition in production matches the adult processing hierarchy reported by Keenan and Hawkins (Figure 8.3). The same order of acquisition was found in German children.[3] Diessel and Tomasello argued that aspects of their results were not consistent with filler-gap distance processing accounts. For example, dependency-locality would predict that S- and A-relatives are equally difficult to process, contrary to what they found in children. Moreover, dependency-locality would

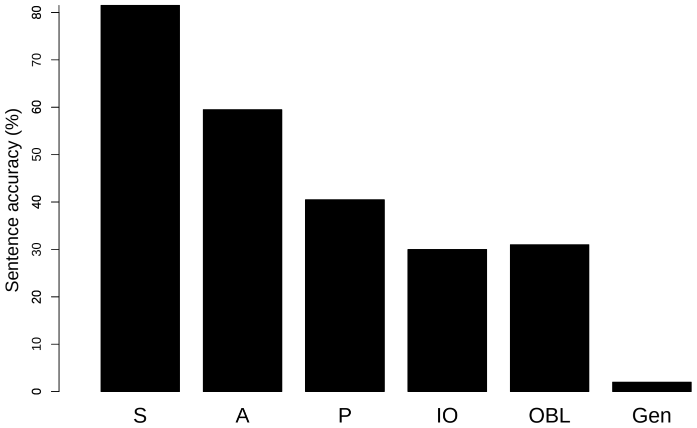


Figure 8.3: The order of relative clause acquisition in English children measured in production (Diessel and Tomasello, 2005).

[3] Notice that unlike Keenan and Hawkins (1987), Diessel and Tomasello (2005) distinguish subject-relatives into intransitive (S-relatives) and transitive relative clauses (A-relatives).

predict that IO-relatives are more difficult than OBL-relatives but Diessel and Tomasello did not find a significant difference in acquisition. To explain their data, they instead proposed an account where frequency of structures and similarity between structures in the primary linguistic data were responsible for creating the hierarchy in development. On their account, subject-relatives (S + A) are easier than P-relatives, because in these constructions the head noun expresses the actor of the relative clause just like the sentence-initial NP in simple transitive clauses. Due to this similarity, S- and A-relatives benefit most from the high frequency of transitive single-clause sentences in child-directed speech. Performance on P-, IO- and OBL-relatives is comparable, according to Diessel and Tomasello, because they share the same subsequence of word categories (NP THAT NP VERB). OBL- and IO-relatives are more difficult because compared to P-relatives they are highly infrequent in the input.

8.3 The recursive Dual-path model approach

In my own approach, I aimed at testing the validity of aspects of the Diessel and Tomasello (2005) account within a computational model of relative clause acquisition. This learning model is exposed to message-sentence pairs from an artificial, English-like language, and it displays differential performance which matches the relative clause hierarchy in development. If the AH behavior is due to the nature of the model's language learning algorithm, then we should be able to manipulate and even remove the hierarchy just by changing the input.

Experimental work cannot substantially manipulate the nature of the input children receive. Hence, it is difficult to assess the degree to which the input shapes syntactic behavior in development. A model of relative clause acquisition allows us to change the input over development, measure the effect of these changes in the model's behavior, and derive empirical predictions for humans. I will focus on three aspects of the model's account of syntax acquisition. Since the model uses an incremental learning mechanism, its syntactic representations develop slowly and are sensitive to the frequency of subsequences of syntactic categories in the input (e.g., THAT ARTICLE NOUN). If we manipulate the input distribution and find that the AH behavior is changed, this would suggest that substructure frequencies may play a part in the construction of the hierarchy. Another feature of the model is that it learns syntactic alternations in which two distinct surface structures are associated with a similar meaning (e.g., active transitives `the man chased the dog` and passive transitives `the dog was chased by the man`) and these structures interfere with each other. We can determine whether this interference is related to the AH behavior by having the model learn a language without alternations. Thus, a computational approach offers three advantages over developmental studies with children. First, we can manipulate the frequency distribution of the input and thereby derive predictions about the composition of children's language input in acquisition. Secondly, we can systematically remove constructions from the input and thereby trace patterns of interference between constructions. And third, we can change the semantic representations of these constructions to determine the extent to which semantic similarities shape the AH in development. By examining how frequency, interference, and meaning relate within a particular account of syntax acquisition, we might be able to make more explicit how language universals like the accessibility hierarchy are influenced by the input.

The specific model of syntax acquisition which I was using is the Dual-path sentence production model (Chang et al., 2006). This connectionist system is built from a simple-recurrent network (Elman, 1990) augmented with a second processing pathway in which the sentence message is represented for production (see Figure 3.2, page 56). It learns the syntax of a target language by mapping meaning representations (input) onto appropriate sentence forms (output). For the current task I extended this model to accommodate processing of multi-clause utterances. The message input to the model used three components, thematic roles (Agent, Patient, Recipient, etc.) which were represented at the where-layer, concepts (lexical semantics) which were represented at the what-layer, and event features (e.g., the number and relative prominence of participants) which could be activated in the event semantics-layer. To encode a message at the beginning of production, thematic roles in the where-layer were temporarily bound to concepts in the what-layer, and the appropriate features in the event semantics-layer were activated. Most importantly, I added information about the co-reference of participants in different events to the message representation of the Chang, Dell, and Bock (2006) model. From this information the model learned to omit the correct participant in the relative clause event when different event roles competed for relativization. For example, the message for A-relatives (`the man that chases the dog`) contained a feature which linked the head noun `the man` in the main clause event to the transitive agent of the subordinate clause event. In a P-relative (`the man that the dog chases`), on the other hand, a feature bound `the man` to the patient role in the relative clause. In this way the model could semantically distinguish similar transitive events and learn to map the corresponding messages onto the correct sentence forms (A- versus P-relatives).[4]

8.4 Language and method

The language I used to train the model contained the basic structures needed to reproduce the relative clause hierarchy in acquisition, and included the transitive and ditransitive alternations (table 8.2). Diessel and Tomasello (2005) argued that children of the tested age might have difficulties with relative clause constructions containing a main clause which expresses a full proposition, and in particular with center-embedded constructions. Therefore, similar to the test items in the Diessel and Tomasello study, multi-clause constructions which the model was exposed to had a relative clause at-

[4] For details regarding the model's architecture and message representation, see Chapters 3 and 4.

Structure	Example
Presentational	`there is a boy .`
Transitive	`the woman kick -ed the teacher .`
Transitive passive	`the teacher was kick -ed by the woman .`
Intransitive	`the cat was sleep -ing .`
Prepositional dative	`a girl throw -s the stick to the cat .`
Double object dative	`a girl throw -s the cat the stick .`
Oblique	`the nurse is play -ing with a dog .`
Relative clauses	`there is the boy that the woman chase -s .`
	`there is a woman that throw -s a cat the toy .`
	`there is a man that the dog was run -ing with .`

Table 8.2: Basic construction types in the language to train the recursive Dual-path model.

tached to the predicate nominal of a presentational clause.[5] Relative clauses which were assembled from presentationals and the structures of Table 8.2 could have all participant roles relativized. The head noun of dative constructions, for example, could be the agent, theme or recipient of the relative clause. The input grammar had verb tense and aspect, inflectional morphemes were represented as separate lexical items. The lexicon contained 56 words in 14 categories which allowed the creation of roughly 2.4×10^6 different sentences. The model was trained on a set of 10.000 sentences from this input language, and tested periodically on 500 novel sentences after every 1.000 training items. The test sentences were randomly generated from the five sentence types which were used in the Diessel and Tomasello experiment.

8.5 The accessibility hierarchy in the recursive Dual-path model

With this input language and training conditions, I replicated the relative clause hierarchy in the recursive Dual-path model (Figure 8.4). Figure 8.4 and all subsequent graphs show the results of averaging performance over 20 model subjects. Model subjects differed with respect to the randomly generated training set to which they were exposed. The x-axis represents the number of sentences that the model has been trained on, the y-axis represents the sentence accuracy for the five tested constructions. Sentence accuracy was measured in terms of perfect match, ignoring minor errors such as wrong determiners, verb tense and aspect. This measure is similar to how Diessel and Tomasello evaluated children's errors in their experiments. At the end of training, the model reached an adult state where it could accurately produce all of the sentence structures.

[5] The Dual-path model is sufficiently general to allow the processing of utterances with full main clauses and multiple embeddings, see Chapter 6. I therefore refer to it as recursive Dual-path model.

We observe that relative clause constructions develop in the same order in the model as in children according to the Diessel and Tomasello (2005) study.

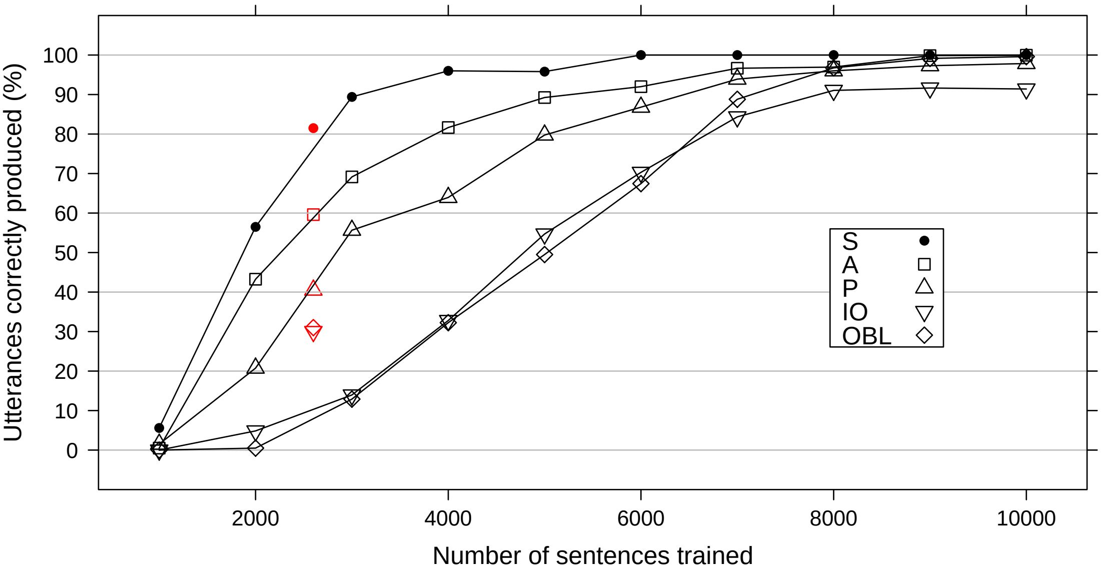


Figure 8.4: The order of relative clause acquisition in the recursive Dual-path model corresponded to the positions on the accessibility hierarchy (Diessel and Tomasello data superposed at epoch 2.500).

To explore what role the input played in creating the hierarchy, I manipulated the model's input, but used the same test set throughout. Therefore, the filler-gap distances remained the same across input manipulations. A processing account would predict that the AH should be robust over small changes in the input. If it is possible, however, to change or remove aspects of the hierarchy, that would suggest that the input might play a larger role in the development of the hierarchy than previously thought.

8.5.1 The S>A contrast

First, I focused on the contrast between S- and A-relatives in a model which was trained on the full language. In the AH condition, S- and A-relatives differed on several features such as length, frequency, binding information, and participation in alternations. If we can determine which of these features are important in the model's S>A behavior, that might indicate how the human syntax acquisition system could be influenced by these factors. Input in the hierarchy condition of Figure 8.4 made several assumptions about the frequency of different structures. For example, S-relatives were more frequent than A-, P-, and IO-relatives, which all had the same frequency, and OBL-relatives were least frequent. To see how those assumptions influenced the model's S/A difference, I equated the frequency of structures in the learning phase (Figure 8.5). For equal frequencies, the

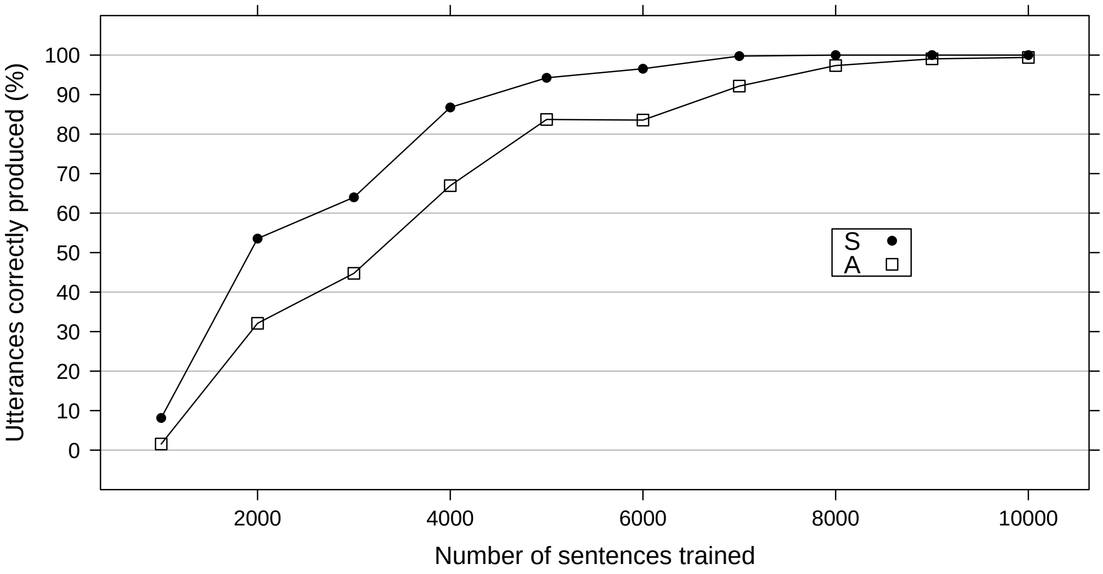


Figure 8.5: The S>A difference persisted when the frequency of all tested constructions was balanced in the input.

accuracy of S- and A-relatives decreased compared to the AH condition, but S-relatives were still learned significantly faster than A-relatives.

Another difference between S- and A-relatives was their overall length. The recursive Dual-path model contains a simple recurrent subnetwork, which is sensitive to the length of input sequences. Thus, we might expect that the S>A difference could be due to the different length of these two constructions. To examine this possibility, I balanced the sentence length in the five test structures by appending prepositional phrases, e.g.,

(2) `there is the man that run -s in the park at night .` (S-relative)

(3) `there is a man that chase -s a dog down the hill .` (A-relative)

When sentence length was balanced, I found a pattern similar to the conditions of Figures 8.4 and 8.5, except that the learning of both structures was delayed (Figure 8.6). Neither input manipulation erased the difference between S- and A-relatives in the model, which suggests that this difference is not due to overall sentence length or input frequency.

A third difference between the two structures lies in the meaning information they require. A- and P-relatives differ in terms of the position of their gap. Therefore, to be able to produce these structures correctly, there had to be a feature that marked the gapped element in the message. Without this information, the model cannot determine whether to produce an A- or a P-relative. S-relatives, on the other hand, do not create this ambiguity, since there is only one possible role to relativize. Hence, part of the S>A difference may have been due to the dependence of the A-relative on meaning

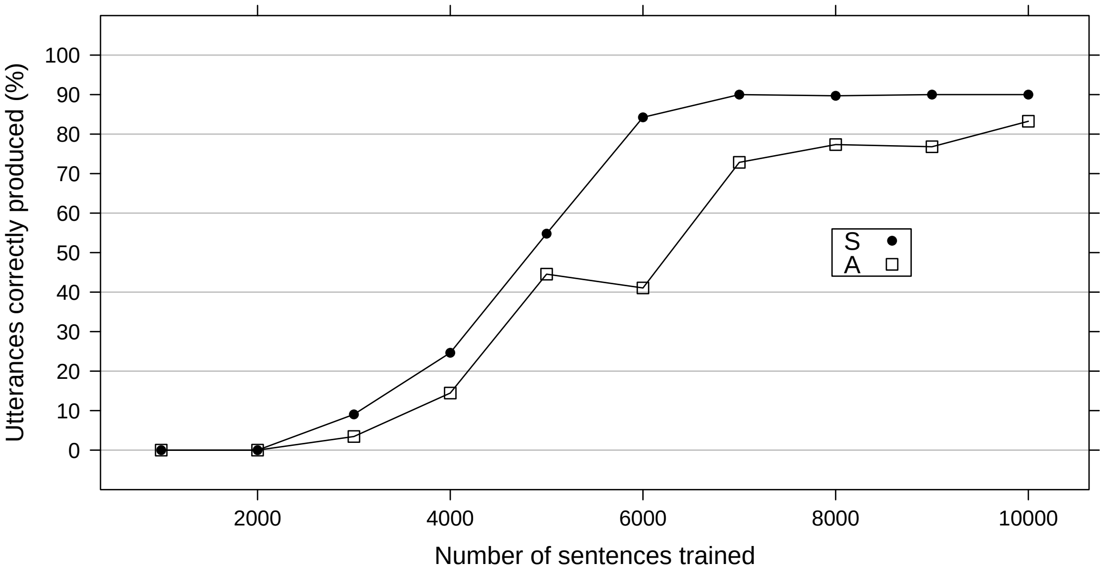


Figure 8.6: Equating the length of all tested constructions did not erase the S>A difference in the model's syntactic development.

information in the message. To examine how much these constructions depended on meaning information, I ran a simulation without role and co-reference information in the event semantics. In this condition, the model received no information about the number and relative prominence of event participants and no role-binding information which would indicate the specific relative clause structure intended for production. As shown in Figure 8.7, this model had trouble learning most of the constructions, except for the S-relatives which were still learned to an adult degree. This suggests that the model found it easier to produce messages which were unambiguously associated with one structure versus those which competed with other structures in the language (like A- and P-relatives). Furthermore, since relative clause acquisition also matched the hierarchy in this condition, the hierarchy did not seem to depend on the particular message representation used in the recursive Dual-path model.

The production accuracy of S-relatives was insensitive to the message manipulation. To demonstrate that the input is critical for explaining the S>A difference, we would like to be able to remove this difference by just manipulating properties of the input. Since the S>A difference was robust over changes in the meaning, and when length and frequency were equated, a more radical manipulation of the input was required. First, I reduced the frequency of S-relatives in the input to half the frequency of A-relatives. Events described by A-relatives have twice as many participants as events described by S-relatives. Because the model is learning to sequence participant roles at the WHERE-layer and more roles entail more distinct sequences, the number of roles might be a critical factor. Balancing S- and A-frequency in the described way controlled for this

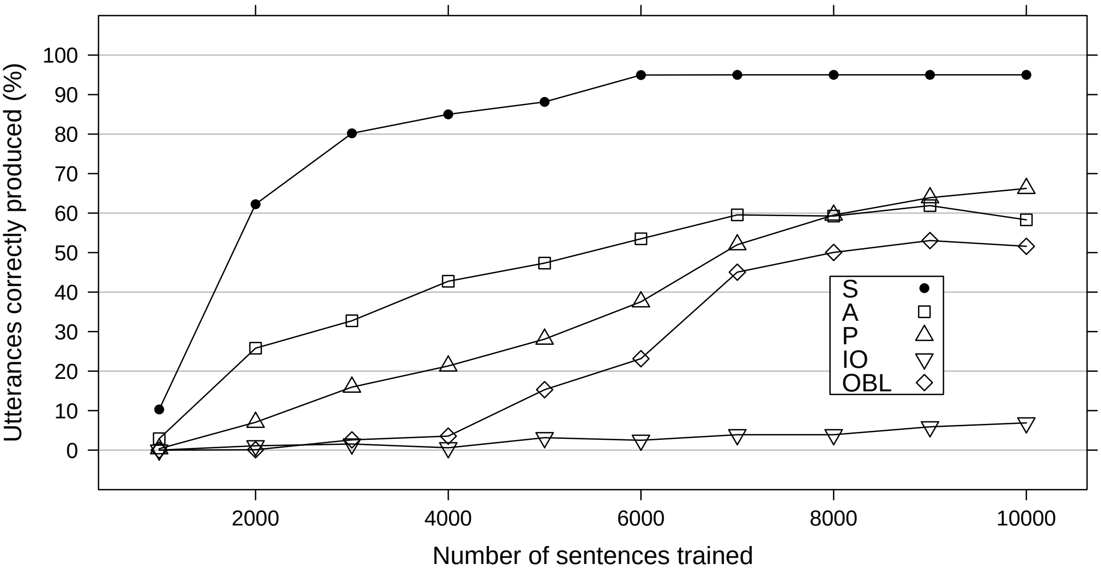


Figure 8.7: Removing participant roles and binding information from the message did not eliminate the S>A difference in the model.

difference. And secondly, I removed input structures that make A-relatives difficult to learn, namely passive transitives. Passive transitives complicate the meaning-to-form mapping the model has to acquire in that they invert the sequence of event participants in the surface form of active transitives.[6] Thus passives might interfere with learning A- and P-relatives. S-relatives, however, do not participate in alternations in which similar messages are mapped onto different sentence forms, and this distinction between S- and A-relatives might partially explain the S>A difference. When both factors were combined the model learned A-relatives as fast as S-relatives (Figure 8.8). Hence, even though the model displayed a strong bias towards S-relatives over all other structures in the hierarchy, this bias could be erased by manipulating the model's input distribution. This demonstrates that the types of structures and their frequencies in the input are crucial for the S>A contrast in the accessibility hierarchy, and suggests that the S>A difference in development may not be maintained in a learning system if the input does not also support that difference.

To summarize the S/A-contrast, frequency and length did not explain why S-relatives were learnt faster than A-relatives. S-relatives were easier for the model because they did not participate in alternations and did not have a main/relative clause binding ambiguity. Thus, the S>A difference was due to inherent factors, like the number of roles, but also due to the learning problem posed by the existence of multiple ways of conveying the same meaning, as in the active/passive transitive alternation.

[6] See also the discussion of alternations in Chapter 5.

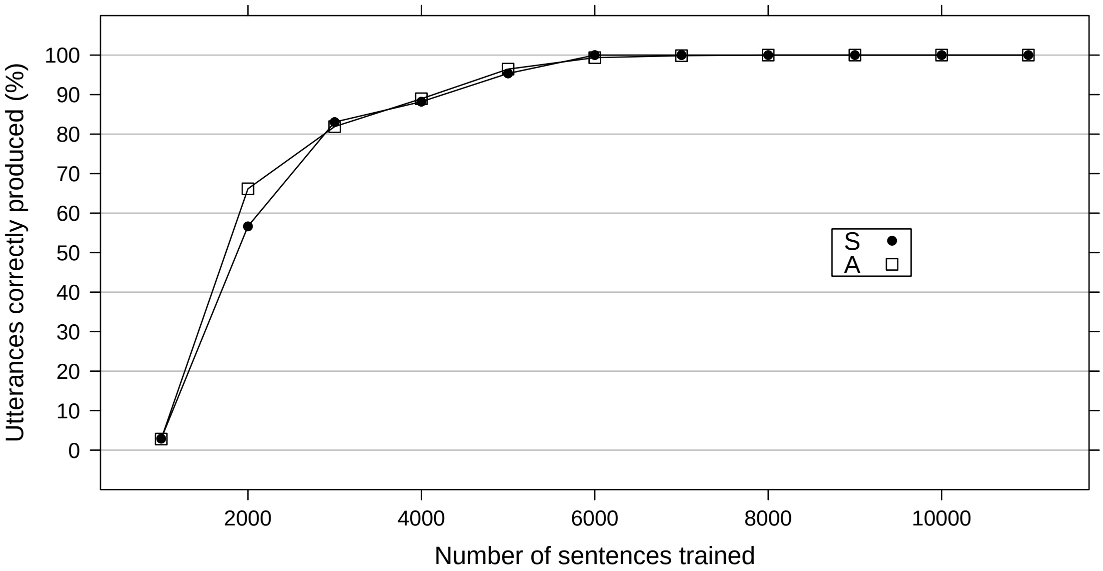


Figure 8.8: S-relatives equaled A-relatives when S-frequency was reduced and passive transitives were removed from the input language.

8.5.2 The A>P contrast

In the hierarchy condition (Figure 8.4, page 261) we observed that the model performed significantly better on A-relatives than on P-relatives despite both structures occurring with the same frequency in the input. This behavior is in line with a large number of comprehension studies which found that object-relativized structures are harder to process than subject-relativized structures across languages, both for adults and children (Hakes, Evans, and Brannon, 1976; Wanner and Maratsos, 1978; Keenan and Hawkins, 1987; King and Just, 1991; Villiers, Flusberg, Hakuta, and Cohen, 1979; Friedmann and Novogrodsky, 2004; Frazier and Clifton, 1989; Gordon, Hendrick, and Johnson, 2001; Tavakolian, 1981). Processing accounts such as Just and Carpenter (1992) and Gibson (1998) argued that this asymmetry was due to a processing bias against object-relativized structures which require more cognitive resources than subject-relativized structures.

Diessel and Tomasello (2005) suggested an alternative account of the A>P difference based on the surface sequence of syntactic categories. A-relatives contain the subsequence THAT VERB, whereas P-relatives contain the subsequence THAT ARTICLE NOUN. Since all of the relative clause structures can relativize subjects, THAT VERB substructures might be more common in a learner's environment than THAT ARTICLE NOUN substructures. If speakers are sensitive to the frequency of substructures, this could help explain the A>P difference. To explore how substructure frequencies related to the A>P difference, I manipulated these frequencies in the model. The model should be sensitive to substructures, because it used a simple-recurrent sequencing network which learned statistical relationships between sequences of adjacent syntactic categories (Elman, 1990; Chang, 2002).

The difference between relevant substructure frequencies can be levelled out, for instance, by manipulating the relativization ratios in dative structures (prepositional dative and ditransitive) in the training set. I reduced the frequency of THAT VERB by reducing the frequency of subject-relativized datives (`there is a man that give -s the toy to the dog`) and increased the frequency of THAT ARTICLE NOUN by increasing the frequency of object-relativized datives (`there is the toy that a man give -s the dog`). Manipulating dative frequencies allowed me to leave the transitive A- and P-frequencies invariant. As a consequence, I was able to remove the A>P difference in development (Figure 8.9). This input manipulation demonstrates that it was not the frequency of the whole construction which was critical, but rather the frequencies of shared substructures that determined the A>P difference in the order of acquisition.

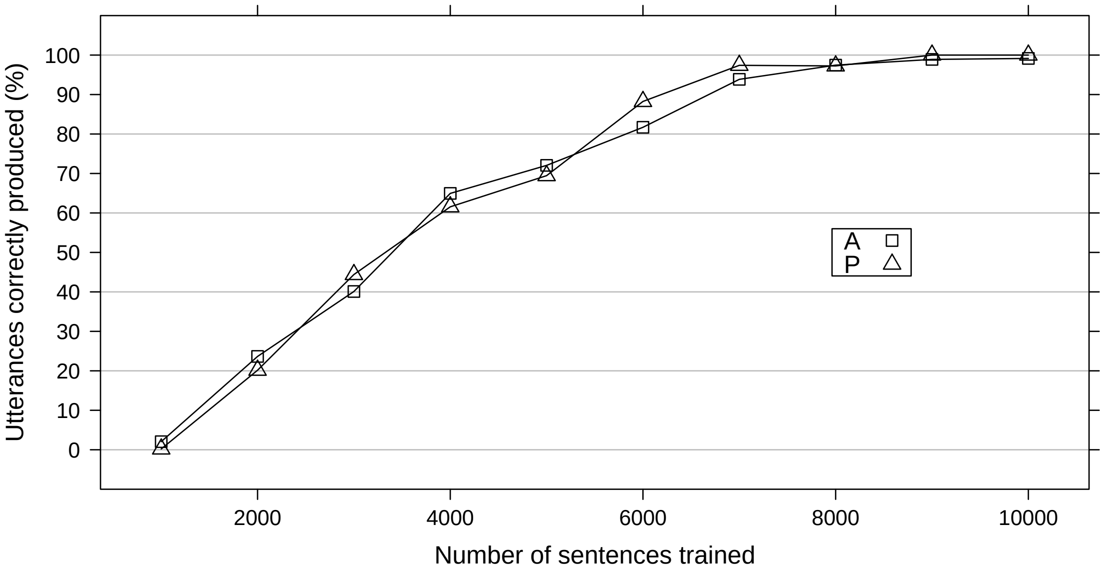


Figure 8.9: A-relatives equaled P-relatives when substructure frequencies were balanced by adjusting the dative relativization ratios while leaving the transitive frequencies intact.

If this account is correct, we can predict that THAT VERB substructures should be more frequent than THAT ARTICLE NOUN in the input to English speaking children. In an analysis of the mother's speech in a dense English corpus (Maslen, Theakston, Lieven, and Tomasello, 2004), I found 157 examples of ARTICLE WORD THAT VERB (where VERB comprised only verbs morphologically marked by -ed or -es). But when searching for cases like ARTICLE WORD THAT ARTICLE, I found only 67 instances. Hence, even without auxiliaries and plural agreement, THAT VERB is far more common than THAT ARTICLE NOUN. This provides support for the explanation of the A>P difference in terms of substructure similarities and indicates that the model can be useful in determining what kinds of units to search for in a corpus analysis. I therefore suggest, that (in analogy to the model) the A>P difference in children could be due to the relative frequencies of common substructures in all of the sentences in the input, rather than reflect a universal processing bias against object-relativized structures.

8.5.3 P-, IO-, and OBL-relatives

The performance differences for P-, IO- and OBL-relatives can be similarly reduced or even inverted by changing the model's input distribution. Each of these constructions was influenced by several *distinct* factors in complex ways. Since these constructions were not significantly different from each other in the Diessel and Tomasello (2005) data, I only report the factors which seemed to have the strongest effect on each construction in the model. The learnability of P-relatives was influenced by many of the factors I mentioned in earlier sections, but in addition, P-relatives were also strongly influenced by the frequency of subject-relativized passives (e.g., `there is a man that was chase -ed by a dog`) which are in direct structural competition. Although these structures are infrequent in child-directed speech, children must hear them or related structures in order to acquire an adult grammar. For statistical learning systems like the recursive Dual-path model, low frequency can be detrimental. In the hierarchy condition (Figure 8.4, page 261) passives were a total of 4.8% in the input. To model the interference effect of learning embedded passives on the development of P-relatives in children, I had to increase their frequency in the model's input. I found that increasing the frequency of subject-relativized passives significantly reduced the accuracy of P-relatives. This effect could further be amplified when I made active and passive transitives less distinct in their message representation. This was implemented by reducing the difference in activation between role features which marked the relative prominence of participants in the event semantics. The result of this manipulation is shown in Figure 8.10 (top) after training on 5000 sentences. P-relatives in this condition went down to the accuracy level of IO-relatives and OBL-relatives in the hierarchy condition of Figure 8.4.

As with the P-relatives, IO-relatives were sensitive to demands of mapping similar messages onto two structures (the dative alternation). In other words, the ditransitive construction (`there is the dog that the girl give -ed a toy`) complicated the acquisition of IO-relatives in the model in a similar way as passive transitives complicate the learning of A- and P-relatives. By removing the ditransitive from the input language, I increased the accuracy of IO-relatives to the level of P-relatives in the hierarchy condition (Figure 8.10, middle).

The OBL-relative construction, on the other hand, was most sensitive to frequency because it was not in direct competition with other input structures. However, OBL-relatives shared semantic similarities with S-relatives, because in my input language oblique objects were treated as prepositional complements of intransitive clauses. Therefore, I expected them to be easily learnable in the model when frequencies of constructions were equal, and this was indeed the case (Figure 8.10, bottom). Hence, the model's account of the low OBL-relative accuracy in the hierarchy condition required

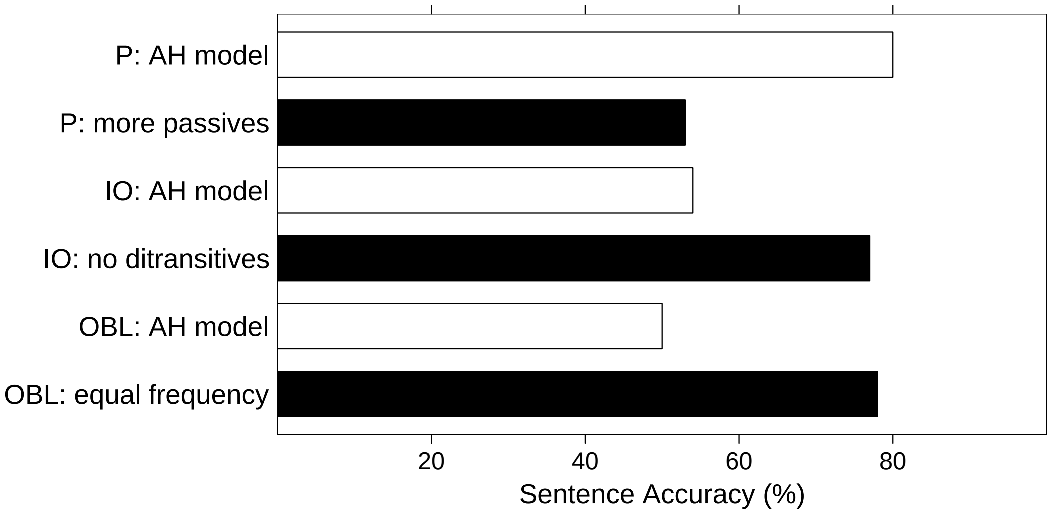


Figure 8.10: Distinct factors influenced the learnability of P-, IO-, and OBL-relatives in the model after training on 5.000 items.

that these structures were much less frequent in the input than S-relatives. Support for this account comes from a corpus study by Diessel (2004) who found that out of all of the relative clauses in a corpus of child-directed speech, 35.6% were S/A-relatives, but only 7.6% were OBL-relatives.

8.6 Eliminating the relative clause hierarchy

If filler-gap distances are not crucial for creating the hierarchy, we should be able to find an input condition in which the model learns a language that does not display the AH in development. I achieved this by creating an input environment which only contained single-clause utterances (e.g., `the dog chase -s the cat`) and sentence tokens of the five tested structures (i.e., S-, A-, P-, IO-, and OBL-relatives) in training. This manipulation removed any interference effect of syntactic alternations (passive transitive and ditransitive) in the model's performance on the hierarchy structures. Secondly, this limited the relativization possibilities for some tested constructions in that subject-relativized obliques and subject- and theme-relativized prepositional datives were removed from the input. In Section 8.5.1 we saw that the number of participant roles in the embedded clause had some influence on the S>A contrast. If we want to eliminate the hierarchy we therefore need to equate for the number of roles in different constructions. Thus, I made the frequency of each relative clause construction in the input proportional to the number of its roles. That is, for each construction, input frequency divided by the number of its roles was identical. In this condition, the hierarchy disappeared (Figure 8.11). This experiment shows that I controlled all the relevant factors that influenced the hierarchy over development in the model. When only the tested structures from the

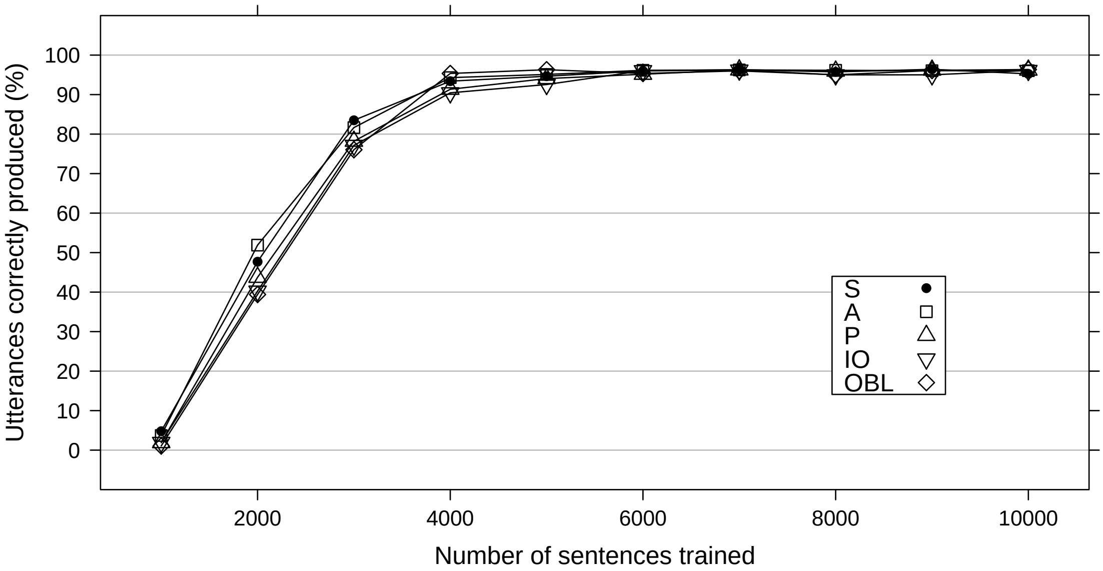


Figure 8.11: When the input language did not contain alternations, and no structures with competing roles relativized, the hierarchy was erased.

accessibility hierarchy were in the input, the same model which previously matched the order of relative clause acquisition in children (Figure 8.4) now behaved entirely neutral with respect to the different sentence structures.

The stepwise elimination of the accessibility hierarchy suggests that patterns of interference and facilitation between the tested items and constructions in the language *outside* the test set brought about the hierarchy in development. Processing theories attempt to define some universal metric rooted in notions of syntactic complexity to determine the processing difficulty of relative clause structures. Here it was shown that the processing difficulty of individual constructions crucially depended on the rest of the input language to a learning mechanism. Hence, the processing difficulty of a construction can not be measured in isolation from the linguistic environment; seeking to identify a complexity metric might be a futile endeavor. I argued instead that it was the diversity of the total input language as filtered through the architecture of a model of syntax acquisition which made some structures harder than others.

The proposed account of the relative clause hierarchy is quite radical. While processing theories largely ignore the input in their explanation, and usage-based accounts of syntax acquisition explain some aspects of relative clause development in terms of the input, I argued that the entire accessibility hierarchy can be reduced to properties of the input language within the framework of a computational learning model.

8.7 Discussion

I showed that a neural network model of syntax acquisition and sentence production was able to exhibit evidence of the AH in syntactic development when given English-like input. However, when that input language was distorted, such that it no longer resembled a natural language, the model's AH behavior was also distorted. I argued that universal properties of natural languages, such as the existence of structural alternations, similarity in meaning between different constructions, and consistent frequency across different languages, may play a part in making the AH a universal feature of human languages.

In addition to providing an account for AH behavior in development, the model suggests how the mechanisms proposed in experimental work (Diessel and Tomasello, 2005; Brandt, Diessel, and Tomasello, 2008) might be implemented. For example, Diessel and Tomasello explained structural errors in their data by stipulating that S/A-relatives are easier to activate than other structures. The model suggests that the frequency of THAT VERB over THAT ARTICLE NOUN across all of the constructions in the language was partially responsible for the ease of activating S/A-relatives. These substructure representations were learned, because the model's simple-recurrent sequencing network attended to local statistical regularities.

The model not only implements mechanisms that have been proposed in the literature, but also emphasizes factors in the AH that have not been considered important for relative clause acquisition. One such factor is syntactic alternations. The model was designed to map from meaning to forms and to handle syntactic alternations, which were therefore included in the language input. Likewise, children have to learn the transitive and dative alternations to become adult speakers. But what I found was that alternations increased the competition between roles for relativization and tended to complicate the generation of forms. This seemed to be important for explaining developmental patterns for different constructions. Along the hierarchy positions, more roles entailed more alternation options; none for intransitives, one for transitives (passives), and two in datives (passive ditransitives). My results suggest that structural diversity incurred from syntactic alternations could be a crucial factor of differential relative clause development in children. Therefore, experimental work on the AH might profit from looking at the influence of alternations.

Accounts of the universal nature of the AH have focused on processing difficulty as the driving force behind the hierarchy. But the presented work with the recursive Dual-path model, which is a sentence processor with a limited capacity memory, indicates that the AH is not an inevitable consequence of sentence processing. No matter how complex a structure is, a model which learns its syntactic representations can recode this structure in a way that requires a minimal amount of memory. This suggests that the learning mechanism may play an important role in determining the complexity of syntactic representations.

More generally, this work suggests that processing accounts are not the only way to explain the universal nature of the AH. Within a computational learning model, I identified properties of the input language, such as substructure frequency, semantic similarities among relative clause types, and the presence of syntactic alternations, which could lead to the hierarchy in development. If these properties are universal across languages, then this learning account provides an alternative explanation of the hierarchy's universality. Any language which violates at least one of these three properties would be an interesting test case for this approach to the accessibility hierarchy. Such a test case, however, may not exist in the pool of natural languages.

Chapter 9

Conclusions

Language is the hallmark of cognition, and complex syntax might separate human from animal communication. Complex syntax is believed by many to be out of reach of neural network models. The present work argued that this might be a premature verdict, if we assume that language processing involves computing with meaning. Instead of recapitulating the results of this thesis in detail, I want to discuss in a more general fashion what I believe are some key ideas which were explored here.

9.1 Key findings

9.1.1 Learning as transduction

It was a working assumption central to this research that natural language is learnable from positive input by means of general cognitive abilities. The basic units of language which must be learned and generalized are constructions—pairings of meaning and form. On this view, syntax is not an autonomous device that needs to be acquired on top of semantic or phonological devices, with interfaces between them that need to be explained (such as syntax-semantics linking rules). Syntax arises in the learning system because it is the very substrate that enables the system to successfully map between meaning and form. Consequently, syntax is not learned prior to or independently of semantics but it is derivative of learning to perform meaning-to-form transduction. The transduction device consists of a mechanism for mapping constructional meaning to grammatical sequences of word categories and a mechanism for mapping sentence-specific content to word forms. Moreover, syntax is not induced from meaningless word sequences and their distributional properties. The induction paradigm in which the learning target is to reconstruct grammars from a set of observations might be misguided as a model of syntactic development in children. In the Dual-path model, syntax forms automatically, in a self-organizing manner, on the basis of domain-general learning procedures and driven by mismatches between internal predictions and overheard utterances. It is not the primary learning target but a byproduct of a more general task:

to make sense of the ambient language, and to be understood in speech. Knowledge of syntax is implicit and resides in the very processor itself rather than a knowledge base external to the processor. The mechanisms by which syntactic knowledge develops in acquisition continue to function in essentially the same way throughout adulthood.

Although the Dual-path model is a model of sentence production, it essentially is a model of the relationship between meaning and word sequences and the learnable mappings between them. The idea that language acquisition primarily involves learning such mappings has not been explored systematically in a computational setting, and this work provides a first step in this direction.

9.1.2 Generalization with semantic similarities

It was shown that the Dual-path model could generalize in quite remarkable ways on a number of tasks. For instance, it could combine familiar constructions into structures with a novel hierarchical organization, and it displayed strong semantic systematicity, the capacity to generalize familiar lexical items to novel roles across embeddings. It also displayed recursive productivity in that it could produce grammatical sentences with three and four nested relative clauses without prior experience, and it could produce correct complex polar interrogatives without positive evidence of these structures in the learning environment. These feats were accomplished despite exposure to only a small number of sentences, typically around 10.000, from artificial languages that could generate up to 4.8×10^{22} different tokens.

The model could achieve this because the complexity of the learning task could be decomposed into simpler subtasks. The large number of sentence tokens generated by the language grouped into a few hundred (or, in some cases, a few thousand) basic construction types. These constructions had a distinct underlying meaning, signalled to the model by the event semantics. From these representations of constructional semantics, the model could learn to activate sentence-specific content in the meaning system in the right sequence. Once lexical meaning was learned, the model could produce correct words in the appropriate slots, and the sequencing system enforced constraints on word class and order. But meaning was also partially overlapping for different constructions. Different parts of the message representation controlled different subsequences of sentences, and different constructions shared semantic substructures in their message. These shared meaning components allowed the model to produce entirely novel constructions which could be assembled from novel combinations of familiar semantic substructures. Generalization to a large amount of sentence tokens from sparse input was enabled by sentence-specific bindings in the WHAT-WHERE-system and by semantic structure shared between propositions in the EVENT SEMANTICS-layer.

This constitutes, I believe, a really novel approach to generalization in neural networks and computational learning systems in general. In purely syntactic approaches to language acquisition (e.g., with simple-recurrent networks), generalization is based on distributional properties of word sequences. In contrast to these accounts, generalization in the Dual-path model is enabled by shared properties of the semantic structures underlying word sequences. This approach might have wide applicability; it is likely to work for many languages other than English, because languages tend to reuse similar structures to convey similar meanings. Consequently, what the Dual-path model approach to language acquisition suggests is that the structure of meaning might be a rich and powerful source of information in children's syntactic development.

9.1.3 Differential processing due to input factors

The Dual-path model was mainly applied in generalization tasks, but it could also explain differential learning and processing, for example the difference between right-branching and center-embedded recursion, and the degradation in performance with depth of embedding. Both findings were in line with human processing data. In Chapter 8, I looked at more fine grained data from the development and adult processing of relative clauses in English. It was demonstrated that the model matched the noun phrase accessibility hierarchy in performance. I argued that differential behavior was strongly influenced by properties of the learning environment. The model showed a processing bias towards simpler structures with less participants because this reduced competition for structural selection and relativization. This bias, however, could be erased by manipulating the input distribution to the model. Individual contrasts in differential processing could not be explained by constructional frequency alone, but were due to substructure frequencies in the total input. Balancing these frequencies selectively removed contrasts. When all structures that did not occur in the hierarchy itself were excluded from the learning environment, the processing differences between constructions disappeared. This indicated that patterns of interference and similarity between input structures were responsible for differential learning and processing. These results suggested that it was not the intrinsic syntactic complexity of constructions in conjunction with working memory limitations that made them easy or difficult to learn and process. Rather, it was the distributional composition of the input and the interactions between different structures which had to be learned over the same set of connection weights that affected the model's syntactic development and caused differential behavior.

This might have important implications for theories of learning and processing. To explain the order of relative clause acquisition and differences in adult processing one needs to look at the distributional properties of spoken corpora, and in particular the semantic and sequential similarities and dissimilarities among structures in the language. Like the Dual-path model, the human processor might be sensitive to linguistic patterns which facilitate and encumber individual structures in different ways. As a consequence, it may be futile to account for acquisition and processing differences by applying a complexity metric to structures in isolation. Moreover, what the Dual-path model approach suggests is that it is not required to posit innate language universals to account for seemingly universal linguistic behavior. Multifactorial, language-specific accounts of differential acquisition and processing, in which the input plays a critical role, might have more explanatory force.

9.2 Future directions

In learning, generalization and processing, the Dual-path model behavior has been proven consistent with a large body of data (Chang, 2002, 2008; Chang, Dell, and Bock, 2006; Fitz and Chang, 2008, 2009) and these results provide converging evidence that the model captures important aspects of human learning and processing. Nonetheless, many results reported in this thesis are rather preliminary in nature. Computational models can always be improved, and experiments conducted with more methodological rigor and systematic depth (as the appendix B shows). Apart from that, I want to sketch a number of future research projects which naturally tie in with and extend the present work.

9.2.1 Perspective taking

Biological factors constrain the human language system and linguistic experience influences the way in which we learn and process language. But how do these factors interact? On the dominant view, learning and processing are constrained by biological factors which shape the human language system. On an alternative view, the language system itself is adapted in learning and processing through linguistic experience (MacDonald and Christiansen, 2002; Wells et al., 2008). The Dual-path model provides an ideal platform to investigate this issue in the domain of relative clause processing.

Evidence from many languages indicates that some relative clause types are harder to process than others. According to a popular theory these processing differences result from working memory limitations (King and Just, 1991; Gibson, 1998), which can be viewed as a biological constraint. These accounts, however, do not explain why there are differences between structures that induce the same memory load. Moreover, they do not explain lexical effects on processing, e.g., why transitive object-relatives are facilitated by pronominal relative clause subjects (Reali and Christiansen, 2007a; Kidd et al., 2007) and inanimate head nouns (Traxler et al., 2002). Thus, a growing body of data cannot be accommodated by memory-based approaches. Drawing on results from Chapter 8, it might be possible to develop a novel account in which competition between structures and their relative frequency of occurrence shape the language system during learning. In this account, differential processing is caused by syntactic alternations and modulated by their frequency.

It is a universal feature of natural languages that they allow speakers to take different *perspectives* to describe the same event. In English, transitive events can be expressed in active or passive voice and dative events can be expressed prepositionally or with the ditransitive. The prepositional dative can partially be passivized and the ditransitive can be fully passivized (e.g., *there is the dog that was given the toy by the girl*). These syntactic alternations convey the same meaning but change the order of event participants in the sentence form. Thus, language users often have a structural choice how to express a particular proposition. In general, the more participants there are, the more alternations are available—none for intransitives, one for transitives, and two for datives. In addition, speakers can relativize different event participants in each construction. There is one way to express intransitive events with a relative clause, four ways to express transitive events, and eleven ways to express dative events. The syntax of these structures must be encoded in the human language processor through learning from examples. But the more structural variety there is, the more complicated the correspondence between meaning and sentence form becomes. This might have several consequences for sentence processing:

(a) More competition in grammatical encoding between structures which are suitable to express the same meaning by different sentence forms.

(b) More variety entails less exposure to each individual structure.

(c) Competition and reduced exposure impede the activation of one specific structure in speech.

In Chapter 8 it was argued that similar factors influenced relative clause processing in the Dual-path model, and the results suggested that syntactic alternations could partially account for the differential processing of intransitive, transitive and dative structures. It would be interesting to investigate the validity of this explanation in human learning and processing by means of corpus analysis and behavioral experiments. There is some evidence that perspective taking with alternations might be an important factor in human processing. It was shown, for instance, that passive transitives are preferred with theme-experiencer verbs (e.g., *challenge*) and dispreferred with agent-patient verbs (e.g., *kick*) (Ferreira, 1994). Thus, alternations are not used with equal likelihood for all verb classes. Thematic structure makes a passive construction more preferable to express the proposition in

(1) *The cowboy that the sheriff challenged was drunk.*

This preference should be manifest in terms of frequencies in corpora of spoken language, i.e., in the linguistic environment of a learner. A learner might become biased towards using passives which are in competition with the active object-relative clauses and this might make (1) difficult to process. In other words, the preferred use of passive structures for some verb classes might explain why the corresponding active structure is particularly hard. Similar preferences have been observed for the prepositional dative/ditransitive alternation. In speech, the ditransitive is vastly preferred over the prepositional dative and the animacy of the recipient co-varies with verb type (Bresnan and Nikitina, 2003). Prepositional dative relative clauses are difficult for adults and delayed in syntactic development. This might be explained within a frequency-based learning approach in which there is competition between structural alternatives. More fine-grained distributional differences have been found for the ditransitive in that the lexical class of the recipient co-varies with voice. Active ditransitives most frequently take pronominal recipients, whereas passive ditransitives take definite noun phrase recipients (Goldberg, 2006). Since transitives align with ditransitives in terms of argument structure, this statistical trend might affect the processing of transitive relative clauses.

Speakers' preferences in the use of syntactic alternations give rise to distributional patterns in the environment from which language is learned. These patterns influence the way in which humans encode grammatical structure during development. Frequency, structural competition and linguistic experience might explain performance differences observed in this process. To test this idea, one would first have to conduct a large-scale corpus analysis on the differential use of alternations in relative clauses. Insights from this analysis could then enter into computational experiments with the Dual-path model. Systematic manipulations of the model's learning environment might yield predictions which could be tested in behavioral experiments using sentence repetition and elicited production methods. The short-term priming paradigm could be used to track interference effects between structures. Experiments with varying exposure to syntactic alternations over a longer period of time might elucidate the role of experience in relative clause processing. As outlined in the introduction, in this project a computational model would be used to link corpus data with human linguistic behavior. The model provides an explicit acquisition mechanism that is sensitive to structural competition and distributional properties of the input, and allows us to derive predictions which can be tested in human processing.

9.2.2 Cross-linguistic study

The Dual-path model account of relative clause acquisition in English suggested that it was not intrinsic syntactic complexity, biological constraints on working memory, or innate language universals that explained acquisition and processing data. Rather, the noun phrase hierarchy was brought about by patterns of similarity and interference in the language and by distributional properties of the input. It would be important to validate this explanation in a cross-linguistic study with languages that have different relative clause systems. German relative pronouns, for instance inflect according to gender, case and number and typically German relative clauses are verb-final (in contrast to single-clause sentences). Consider the two relative clause types:

(2) *Da ist der Mann der die Katze jagte.*
There is the-NOM man who-NOM the-ACC cat chased
'There is the man who chased the cat.'

(3) *Da ist der Mann den die Katze jagte.*
There is the-NOM man who-ACC the-NOM cat chased
'There is the man who the cat chased.'

Word order in A-relative (2) and P-relatives (3) is identical and, in contrast to English, does not signal the grammatical role of the head noun in the relative clause. Instead, the role of *der Mann* is marked on the pronoun. This can create ambiguities as in

(4) *Da ist die Frau die die Katze jagte.*

because the pronoun *die* can either be nominative or accusative (A- or P-relative). For these (and other) reasons, we can expect patterns of structural similarity and interference to arise in the input which might be very different from English, and it is difficult to predict the effects of these patterns on learning and processing in the model without further simulations.

Another interesting test case for the Dual-path model approach would be Japanese in which relative clauses are prenominal and pronouns are not used. Case is marked by postpositions and the grammatical role of the head noun in the relative clause must be inferred from the case of the embedded complement:

(5) *Uma-o ketta roba-ga sinda.*
Horse-ACC kicked mule-NOM died
'The mule that kicked the horse died.'

(6) *Uma-ga ketta roba-ga sinda.*
Horse-NOM kicked mule-NOM died
'The mule that the horse kicked died.'

A- and P-relatives differ only with respect to a case particle in their surface form. In addition, complements need not be expressed in Japanese relative clauses which gives rise to many possible interpretations:

(7) *Hon-o katta gakusee.*
Book-ACC bought student
'The student (who) bought a book.'
'The student (from whom) ( ) bought a book.'
'The student (for whom) ( ) bought a book.'

As in German (but for different reasons), surface form is sometimes not useful in determining the interpretation of Japanese relative clauses.[1]

Word order in relative clauses differs pairwise in English, German and Japanese. The semantic features the Dual-path model employs in its message representation, however, are generic and work in each of these languages. The model can learn to map such messages onto any word order, provided that meaning-form mappings are sufficiently systematic. Hence, relative clause processing and development in these languages could be studied in the Dual-path model using similar messages across languages which map to very different sentence forms. Due to these word order differences, the model would acquire distinct meaning-form mappings and, as in English, these mappings might be influenced by language-specific similarities between input structures. Thus, it might be possible to modulate differential performance by systematically manipulating the learning environment in order to show that input factors similar to those isolated in Chapter 8 drive acquisition and processing in languages other than English.

[1] Examples (5) and (6) taken from Ishizuka (2005), (7) from Matsumoto (1999).

9.2.3 Dynamic message

In the configuration studied in this thesis, the Dual-path model learned in situated comprehension in which it was assumed that children can infer the meaning of overheard utterances from visual information. While this might be a realistic scenario for single-clause utterances, it is less realistic for more complex multi-clause utterances, specifically those of Chapter 6. The complete message that was given to the model as input provided more of the conceptual structure of utterances than can reasonably be assumed to be available to children at the onset of processing. Similarly, in production it is unlikely that the entire intended message of a sentence (with several embeddings) is provided by an external planning system before production begins. How speakers construct meaning incrementally before and during production is not well-understood, it is conceivable that to a large extent meaning is constructed *as we speak*, and guided by our own self-monitored speech output. It would therefore be an important continuation of this thesis to *dynamicize* the Dual-path model's message representation both in comprehension and in production. With a dynamic message, sentence meaning would only be activated partially in the beginning and more semantic information would become available to the model during processing. Message components would also selectively be deactivated when they have already been utilized and begin to leave the attentional spotlight. This might place stronger demands on the model's working memory system which would have to retain traces of previously activated message components. Thus, in a dynamic version of the Dual-path model, meaning would be constructed incrementally in a window of attention and gradually fade away in working memory as the model progresses through a sentence in word prediction. Such a model could be built on existing research of visual scene analysis which requires tracking multiple objects and their interrelations (Cavanagh and Alvarez, 2005; Ferreira et al., 2008). In particular, research using the eye-tracking methodology to investigate the mechanisms of thematic role assignment in comprehension might be a fruitful source to develop and motivate this dynamic model (Knoeferle et al., 2005; Knoeferle and Crocker, 2008).

The static message Dual-path model makes strong assumptions about the availability of rich semantic representations in learning and processing, especially in case of multi-clause utterances. The results of this thesis must be evaluated against the plausibility of these assumptions. It should be pointed out, however, that a complete and invariable message might not be computationally optimal to learn meaning-to-form mappings which generalize well and in human ways. It was mentioned before (Chapter 5) that the static message creates a serial order and a timing problem for the model (what to predict next, and when to produce what), which both need to be solved by the processor when learning from message-sentence pairs. If, on the other hand, the message is dynamic and incremental, these tasks can partially be relegated to the attentional mechanism which guides meaning construction in visual processing. Message components are activated only at and around relevant sentence positions and this locally restricts the choice of roles, concepts and words to sequence. In the static model all semantic features are active concurrently which makes any sentence constituent a salient choice for the next slot. Hence, a tight interplay between mechanisms of selective attention and word prediction might greatly facilitate the sequencing of thematic roles and the acquisition of syntactic frames in the model. Secondly, in Chapter 6 it was argued that novel combinations of semantic features in the message were detrimental to generalization. The model did not know which thematic roles to sequence when, because active features in one clause interfered with message components in other clauses, and this explained why production accuracy degraded rapidly with depth in recursive processing. If sentence meaning becomes available to the model sequentially and in smaller chunks, this problem might be avoided. Although an input message in its totality might express a proposition that the model has not experienced before, exposure to segments of semantic components might enable the model to incrementally build novel sentences with novel meanings from chunks of feature combinations it is familiar with. In other words, the problem of how learned meaning-to-form mappings can be extended beyond experience might be solvable in a dynamic model in which globally novel propositions are composed of known semantic substructures that are activated over time rather than in parallel.

It can be speculated that a dynamic Dual-path model might be superior in a number of generalization tasks studied in this thesis. With a fixed message, the model cannot genuinely be recursively productive because it cannot process sentences of unbounded length in any specific set-up. More embeddings require more semantic features and thematic roles and these need to be represented in the model and trained in learning. A dynamic message would allow the model to reuse the same feature set over and over again and therefore remove architectural limitations of the current model. Secondly, it was found that the model would not generalize from exposure to one embedding to two or more embeddings, it only generalized from two embeddings to three and four (Chapter 6). With a dynamic message this might be remedied because the model would not have to develop a policy how to deal with multiple conflicting features in novel propositions. It was also found that the difference in accuracy for right-branching and center-embedded structures in the model was perhaps too small to perfectly match human behavior. A dynamic message which incrementally provides information to the model as it moves over the sentence proposition, highlighting semantic features like a spotlight, might create a larger performance difference by making center-embedded structures harder and right-branching structures easier to learn. First results obtained with a dynamic prototype clearly point in this direction. And finally, generalizing auxiliary fronting from simple to complex polar interrogatives might be facilitated by using a dynamic message because semantic features which distinguish questions from declaratives could be activated locally and temporarily only, and this might ensure the integrity of the embedded clause.

To summarize, it would be desirable to develop a dynamic Dual-path model which is more faithful to the processes by which human learners construct meaning incrementally in comprehension and production. The model version studied in this thesis should be viewed as a useful simplification and approximation of such a model. There is no indication that positive results reported here are an artefact of the fixed message representation. On the contrary, a dynamic model can be expected to display even better learning and generalization capacities.

Appendix A

Dual-path model details

In this appendix a few more technical details about the Dual-path model will be provided which might be interesting to other modellers and helpful to assess the presented results.

A.1 Model specification

The size of the model's layers depended on the artificial language used and the computational requirements of specific learning tasks; it could vary across chapters. Roughly, the model had 40–70 WORD (CWORD) units, 30–60 WHAT (CWHAT) units, 40–100 HIDDEN (CONTEXT) units, and between 10 and 35 COMPRESS (CCOMPRESS) units. The number of EVENT SEMANTICS and WHERE (CWHERE) units depended on the maximum number of clauses within sentences of the artificial language. It varied between 10 and 45 for the EVENT SEMANTICS and between 4 and 20 for the WHERE (CWHERE) layers.

All units in the model used the dot product as their integration function. The net input x_i to unit u_i was computed as the weighted sum $x_i = \sum_{j=1}^{n} w_{ij} \cdot y_j$ of all n output signals y_j which u_i received as input from other units in the network. As their activation function, most units in the model used the standard logistic function where the output y_i of unit u_i is computed as $y_i = \frac{1}{1+e^{-x_i}}$ with the net input $x_i \in \mathbb{R}$ and $y_i \in [0,1]$. The model's output is a categorical variable ranging over the words in the lexicon and we wish to interpret this output as a probability vector for the lexical items. To achieve this, the WORD-layer used the soft-max activation function. Let x_i again be the net input to output unit u_i, then the soft-max function computes the output y_i of unit u_i as $y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$ where x_j is the net input to output unit u_j. Thus, the activation of output unit u_i depended on the strength of the input signal to all other output units. In this way, the soft-max function magnified small differences in net input among the units, i.e., it rewarded the winning word and punished all weaker word outputs. Moreover, it normalized the output values of all units in that they sum to 1.[1] To

[1] Note that for two lexical items this function is equivalent to the standard logistic function.

measure the overall error on the WORD units, the divergence function $\sum_{i=1}^{n} t_i \cdot \log(\frac{t_i}{y_i})$ was used, where y_i is the output of unit u_i and t_i is the target value of this unit at the current word position. Apart from the WORD-layer, soft-max was also used at the CWHERE-layer to make the incremental role assignment more distinct. This reduced the noise at the CWHERE-layer deriving from the activation of several concepts at the CWHAT-layer and facilitated structural selection. The activation of EVENT SEMANTICS units was implemented by untrainable bias weights. The strength of the weights was controlled by the message and determined the activation values in the EVENT SEMANTICS-layer which used linear units. A negative bias was injected into the WHAT and CWHAT units to guarantee that their activation level was low when they received no input. The units in the CONTEXT-, CWHERE2- and CWORD-layers were copy units which received their activation from the previous time-step activation of units in other layers. Incoming connections to these layers were not trainable. The CONTEXT-layer received a copy of the previous activation state of the HIDDEN-layer to create a working memory. The CWHERE2-layer time-averaged the CWHERE and its own previous activation state, creating a buffer which gradually accumulated all previously produced roles. At the CWORD-layer, the activation state of the model's output at the WORD-layer and the external overheard word input were summed, and the layer was normalized.

A.2 Training procedure

At the beginning of training, the weights in the network were randomized with a fixed seed for all experiments. For each message-sentence pair in training, the model's prediction error for each word was collected and weights were updated after every training pattern. Thus, one epoch in the training regime corresponded to one sentence, not the entire training set (i.e., the batch-size was 1). Training was non-incremental throughout. In other words, message-sentence patterns to which the model was exposed were randomly selected from the training set and not presented in a particular order (e.g., ordered by syntactic complexity). Weights were updated by steepest-descent backpropagation of error. The learning rate was chosen in between 0.15 and 0.2 depending on the complexity of the learning task, and it was set to decrease linearly over time to a final value of 0.02. Roughly 25% of the total training time the initial learning rate was used, then for 50% of the time it decreased and it remained constant again for the final 25% of training time. In order to prevent weights in the network from becoming too large, I used weight decay in all simulations (except Chapter 8). In addition to each weight update by backpropagation at time t, the size of the weight was decreased by $\kappa \cdot w_{ij}(t-1)$ where κ was set to 5×10^{-7} and $w_{ij}(t-1)$ was the weight size after the previous update. It has been shown that weight decay improves generalization in feedforward networks (Krogh and Hertz, 1992) and this was also observed in the recurrent Dual-path model. Intuitively, large individual weights can create local specialization by masking the contribution of other weights to finding an optimal mapping and this tendency is balanced by the decay term. In similar vein, some simulations in which generalization was the critical performance measure (e.g., recursive productivity in Chapter 6) used synaptic noise on all connections projecting into the HIDDEN-layer. The noise was sampled from a Gaussian distribution with standard deviation 0.005 centered around 1.0, and it was applied multiplicatively, scaling the weights with the noise term. Synaptic noise was shown to improve generalization in recurrent networks when the error surface is complex (as in language learning), because it allows the network to 'jump' out of local minima more easily (Jim, Giles, and Horne, 1996).

A.3 Simulation environment

All experiments were conducted using version 2.63 of the LENS neural network simulator (Rohde, 1999). The software package ran on an Intel(R) Xeon (TM) 3.06 GHz workstation with 2 Gbyte of RAM in a SuSE v.10.1 Linux environment.

Appendix B

Improved question learning

Since the completion of Chapter 7, I managed to significantly improve on the reported results in the '*wh*-questions' condition. Due to time constraints and the comparative nature of Chapter 7, it was not possible to include these results since it would have required a large amount of re-modelling and error analysis. I will therefore only briefly outline these results in this appendix.

B.1 New data

In the '*wh*-questions' condition, the model received training with simple-clause and relative clause declaratives, simple yes/no questions and complex *wh*-questions. It was tested on novel complex polar questions. Figure B.1 shows the learning curves for all these structures in terms of grammaticality. Simple-clause structures and simple polar questions were learned quickly, followed by relative clause constructions and *wh*-questions. At the end of training, the model produced around 37% grammatical polar questions, although these structures were not experienced in the training phase.

B.2 Analysis

Moreover, the model generalized in desirable ways. Similar to humans, it showed a preference for right-branching over center-embedding and a preference for subject-relativized over object-relativized structures (Figure B.2). Out of all grammatical complex questions which the model produced, roughly two thirds were right-branching and roughly two thirds were subject-relativized. To compare these results with those of Reali and Christiansen (2005), I tested the trained model on 1.000 pairs of grammatical and ungrammatical center-embedded questions (i.e., questions in which the main or embedded clause auxiliary was displaced). The model received message input which was neutral between the two forms. The production output was then compared with both

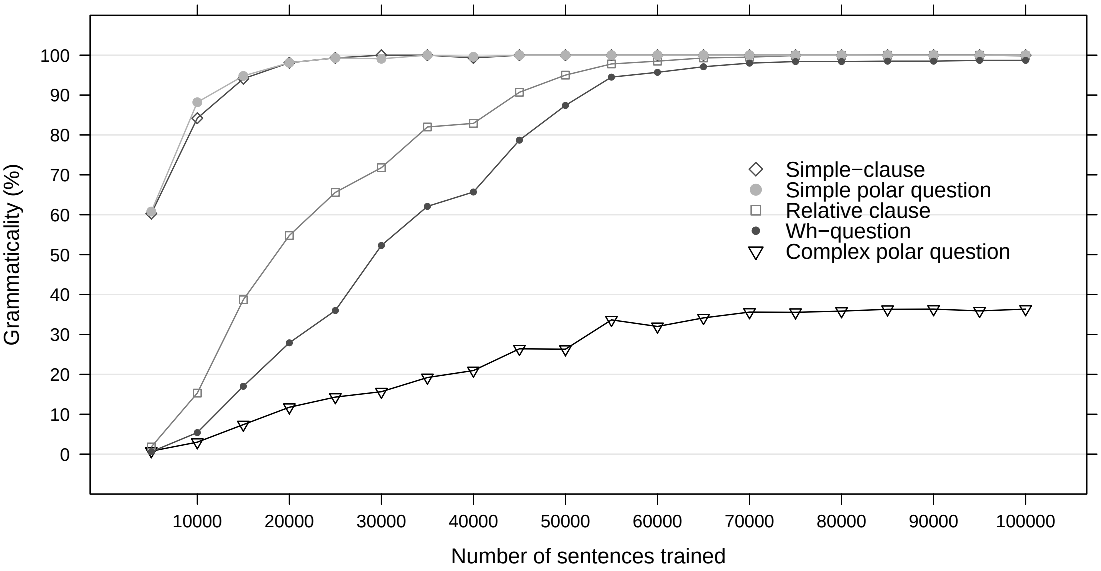
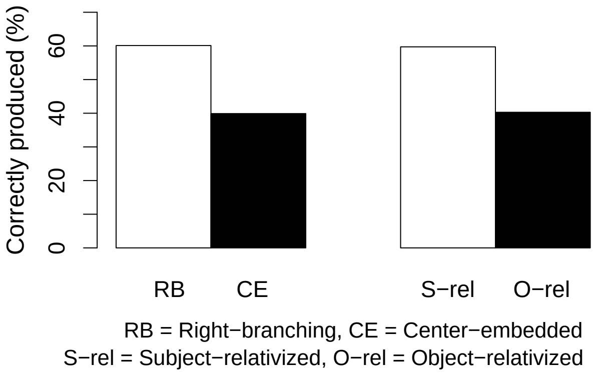


Figure B.1: Complex polar question learning with *wh*-questions in the input.

targets, and classified as either grammatical or ungrammatical based on a graded performance measure (Figure B.3). In 89% of the tested pairs the model's output was closer



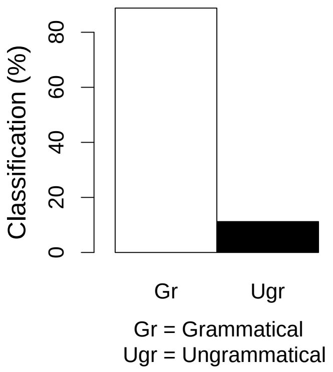


Figure B.2: Preference for CPQRB and subject-relativized questions.

Figure B.3: CPQCE classification after training.

to the grammatical question. Quantitatively these results are similar to those of Reali and Christiansen (2005). My test set, however, contained a large amount of structural variation in the CPQCE and the results did not depend on tagging the model's input in a specific way. Structure-dependent auxiliary fronting was learned by the Dual-path model from semantic overlap with other question types whose occurrence is warranted in child-directed speech.

These new results are described in more detail in Fitz and Chang (2009).

List of Abbreviations

AH	accessibility hierarchy
ACC	accusative case
ATO	active transitive object
ATS	active transitive subject
AUX	auxiliary
BN	binding node
BPTT	backpropagation through time
CE	center-embedding
CFL	context-free language
COMP	complementizer
CP	complementizer phrase
CPQ	complex polar question
CPQCE	center-embedded complex polar question
CPQRB	right-branching complex polar question
CR	compressed representation
CS	cross-serial
CSL	context-sensitive language
DDS	double-object dative subject
DE	depth of embedding
DET	determiner
DFA	deterministic finite automaton
DO	double-object dative
DP	determiner phrase
DPO	double-object dative primary object
DSO	double-object dative secondary object
FGREP	forming global representations with extended backpropagation
FLN	faculty of language in the narrow sense
FSM	finite-state machine
HSD	honestly significant difference
IO	indirect object
IP	inflectional phrase
ITS	intransitive subject
LDA	linear discriminant analysis
LENS	light, efficient network simulator
MCSG	mildly context-sensitive grammar
NP	noun phrase
NOM	nominative case
NVN	noun-verb-noun
OBL	oblique
OBO	oblique object
OBS	oblique subject
OComp	object of comparison
OO	object-modifying, object-relativized

OS	object-modifying, subject-relativized
PAST	past tense
PCA	principle components analysis
PD	prepositional dative
PDP	parallel distributed processing
PDS	prepositional dative subject
PER	period
PP	prepositional phrase
PPO	prepositional dative primary object
PRES	present tense
PROG	progressive aspect
PRON	pronoun
PropN	proper name
PSO	prepositional dative secondary object
PTS	passive transitive subject
PTO	passive transitive object
RAAM	recursive auto-associative memory
RB	right-branching
RC	relative clause
RDR	recursive distributed representation
RNN	recurrent neural network
RT	reaction time
SC	simple clause
SCN	sequential cascaded network
SIMP	simple aspect
SN	subnetwork
SO	subject-modifying, object-relativized
SPEC	subsymbolic parser for embedded clauses
SPQ	simple polar question
SRAAM	sequential recursive auto-associative memory
SRN	simple recurrent network
SS	subject-modifying, subject-relativized
UG	universal grammar
VP	verb phrase
XOR	exclusive or
XYZ	thematic role coding

Bibliography

Allis, V. 1994. Searching for solutions in games and artificial intelligence. Ph.D. thesis, University of Limburg, Maastricht, The Netherlands.

Alquézar, R. and Sanfeliu, A. 1995. An algebraic framework to represent finite state automata in single-layer recurrent neural networks. *Neural Computation 7,* 5, 931–949.

Altmann, G. T. M. and Dienes, Z. 1999. Rule learning by seven-month-old infants and neural networks. *Science 284*, 875.

Altmann, G. T. M. and Kamide, Y. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition 73*, 247–264.

Ambridge, B., Rowland, C. E., and Pine, J. M. 2008. Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science 32*, 222–255.

Arbib, M. A. and Rizzolatti, G. 1997. Neural expectations: A possible evolutionary path from manual skills to language. *Communication and Cognition 29*, 393–424.

Arnold, J., Wasow, T., Losongco, A., and Grinstrom, R. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language 76*, 28–55.

Aslin, R., Saffran, J. R., and Newport, E. L. 1998. Computation of conditional probability statistics by 8-month old infants. *Psychological Science 9*, 321–324.

Bach, E., Brown, C., and Marslen-Wilson, W. 1986. Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes 1,* 4, 249–262.

Baker, C. L. 1978. *Introduction to Generative-Transformational Syntax.* Prentice-Hall, Englewood Cliffs, NJ.

Baker, L. and Wagner, J. L. 1987. Evaluating information for truthfulness: The effects of logical subordination. *Memory and Cognition 15,* 3, 247–255.

Baldwin, D. A. 1993. Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology 29*, 832–843.

Bencini, G. and Goldberg, A. E. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language 43*, 640–651.

Bengio, Y., Simard, P., and Frasconi, P. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks 5,* 2, 157–166.

Berg, G. 1992. Connectionist parser with recursive sentence structure and lexical disambiguation. In *Proceedings of the Tenth National Conference on Artificial Intelligence.* AAAI Press, Menlo Park, CA, 32–37.

Berwick, R. C. and Weinberg, A. S. 1984. *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition.* MIT Press, Cambridge, MA.

Bever, T. G. 1970. The cognitive basis for linguistic structure. In *Cognition and Development of Language*, J. R. Hayes, Ed. Wiley, New York, 279–353.

Blank, D. S., Meeden, L. A., and Marshall, J. B. 1992. Exploring the symbolic/subsymbolic continuum: A case study of RAAM. In *Closing the Gap: Symbolism vs. Connectionism*, J. Dinsmore, Ed. Erlbaum, Hillsdale, NJ.

Blaubergs, M. S. and Braine, M. D. S. 1974. Short-term memory limitations on decoding self-embedded sentences. *Journal of Experimental Psychology 102,* 4, 745–748.

Blumenthal, A. L. and Boakes, R. 1967. Prompted recall of sentences. *Journal of Verbal Learning and Verbal Behavior 6*, 674–676.

Bock, K. 1986. Syntactic persistence in language production. *Cognitive Psychology 18*, 355–387.

Bock, K. and Cutting, J. C. 1992. Regulating mental energy: Performance units in language production. *Journal of Memory and Language 31,* 1, 99–127.

Bock, K. and Loebell, H. 1990. Framing sentences. *Cognition 35*, 1–39.

Bodén, M. and Wiles, J. 2000. Context-free and context-sensitive dynamics in recurrent neural networks. *Connection Science 12,* 3, 197–210.

——— 2002. On learning context free and context sensitive languages. *IEEE Transactions on Neural Networks 13,* 2, 491–493.

Bodén, M., Wiles, J., Tonkes, B., and Blair, A. 1999. Learning to predict a context-free language: Analysis of dynamics in recurrent hidden units. In *Proceedings of the International Conference on Artificial Neural Networks*, D. Willshaw and A. Murray, Eds. Edinburgh, 359–364.

Bohannon, J., MacWhinney, B., and Snow, C. E. 1990. No negative evidence revisited: Beyond learnability or who has to prove what to whom. *Developmental Psychology 26,* 2, 221–226.

Bohannon, J. and Stanowicz, L. 1988. The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology 24,* 5, 684–689.

Boomer, D. S. 1965. Hesitation and grammatical encoding. *Language and Speech 8,*

148–158.

Botvinick, M. and Plaut, D. C. 2004. Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review 111,* 2, 395–429.

Brandt, S., Diessel, H., and Tomasello, M. 2008. The acquisition of German relative clauses: A case study. *Journal of Child Language 35,* 2, 325–348.

Bresnan, J. and Nikitina, T. 2003. On the gradience of the dative alternation. Unpublished draft, Stanford University, `http://www.stanford.edu/~bresnan/bresnan-nikitina.pdf`.

Brink, D. van den and Hagoort, P. 2004. The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. *Journal of Cognitive Neuroscience 16,* 6, 1068–1084.

Brooks, P. and Tomasello, M. 1999. How children constrain their argument structure constructions. *Language 75,* 4, 720–738.

Brown, C. M., Hagoort, P., and Ter Keurs, M. 1999. Electrophysiological signatures of visual lexical processing: Open- and closed-class words. *Journal of Cognitive Neuroscience 11,* 3, 261–281.

Brown, R. W. 1957. Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology 55,* 1, 1–5.

Brown, R. W. and Hanlon, C. 1970. Derivational complexity and the order of acquisition in child speech. In *Cognition and the Development of Language*, J. R. Hayes, Ed. Wiley, New York, 11–54.

Caplan, D., Hildebrandt, H., and Waters, G. S. 1994. Interaction of verb selectional restrictions, noun animacy, and syntactic form in sentence processing. *Language and Cognitive Processes 9*, 549–585.

Carrasco, R. C. and Forcada, M. L. 2001. Finite-state computation in analog neural networks: Steps towards biologically plausible models? In *Emergent Computational Models Based on Neuroscience*, S. Wermter, J. Austin, and D. Willshaw, Eds. Lecture Notes in Computer Science, vol. 2036. Springer, New York, 480–493.

Carrasco, R. C., Forcada, M. L., Valdés-Munõz, M. Á., and Ñeco, R. P. 2000. Stable encoding of finite-state machines in discrete-time recurrent neural nets with sigmoid units. *Neural Computation 12,* 9, 2129–2174.

Castro, J. L., Mantas, C. J., and Benítez, J. M. 2000. Neural networks with a continuous squashing function in the output are universal approximators. *Neural Networks 13,* 6, 561–563.

Cavanagh, P. and Alvarez, G. A. 2005. Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences 9,* 7, 349–354.

Chalmers, D. J. 1990. Syntactic transformations on distributed representations. *Connection Science 2*, 53–62.

Chalup, S. K. and Blair, A. D. 1999. Hill climbing in recurrent neural networks for learning the $a^nb^nc^n$ language. In *Proceedings of the 6th International Conference on Neural Information Processing, Perth*, T. Gedeon et al., Ed. Vol. 2. 508–513.

——— 2003. Incremental training of first order recurrent neural networks to predict a context-sensitive language. *Neural Networks 16,* 7, 955–972.

Chang, F. 2002. Symbolically speaking: A connectionist model of sentence production. *Cognitive Science 26*, 609–651.

——— 2008. Learning to order words: A connectionist model of heavy NP shift and accessibility in Japanese and English. *In preparation.*

Chang, F., Dell, G. S., and Bock, K. 2006. Becoming syntactic. *Psychological Review 113*, 234–272.

Chen, S. F. and Goodman, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language 13*, 359–394.

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

——— 1975. *Reflections on Language.* Pantheon Books, New York.

——— 1986. *Knowledge of Language.* Praeger, New York.

——— 1995. *The Minimalist Program.* MIT Press, Cambridge, MA.

Chomsky, N. and Miller, G. A. 1963. Introduction to the formal analysis of natural languages. In *Handbook of Mathematical Psychology, vol. 2*, R. D. Luce, R. R. Bush, and E. Galanter, Eds. Wiley, New York, 269–321.

Chrisman, L. 1991. Learning recursive distributed representations for holistic computation. *Connection Science 3,* 4, 345–366.

Christiansen, M. H. 1992. The (non)necessity of recursion in natural language processing. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society.* Erlbaum, Hillsdale, NJ, 665–670.

——— 1994. Infinite languages, finite minds: Connectionism, learning and linguistic structure. Ph.D. thesis, University of Edinburgh, Scotland.

Christiansen, M. H. and Chater, N. 1999a. Connectionist natural language processing: The state of the art. *Cognitive Science 23,* 4, 417–437.

——— 1999b. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science 23,* 2, 157–205.

——— 2003. Constituency and recursion in language. In *The Handbook of Brain Theory and Neural Networks*, 2nd ed., M. A. Arbib, Ed. MIT Press, Cambridge, MA, 267–271.

Christiansen, M. H. and Curtin, S. L. 1999a. The power of statistical learning: No need for algebraic rules. In *Proceedings of the Annual Conference of the Cognitive Science Society.* Vol. 21. Erlbaum, Mahwah, NJ, 114–119.

——— 1999b. Transfer of learning: Rule acquisition or statistical learning? *Trends in Cognitive Sciences 3,* 8, 289–290.

Church, K. W. 1980. On memory limitations in natural language processing. M.S.

thesis, Massachussetts Institute of Technology, USA.

Clahsen, H. and Featherston, S. 1999. Antecedent priming at trace positions: Evidence from German scrambling. *Journal of Psycholinguistic Research 28*, 415–437.

Cleeremans, A. 1993. *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing.* MIT Press, Cambridge, MA.

Cleeremans, A., Servan-Schreiber, D., and McClelland, J. L. 1989. Finite state automata and simple recurrent networks. *Neural Computation 1,* 3, 372–381.

Comrie, B. and Keenan, E. L. 1979. Noun phrase accessibility revisited. *Language 55*, 649–664.

Cooreman, A. and Sanford, T. 1996. Focus and syntactic subordination in discourse. Research Paper no. RP-79, University of Edinburgh, HCRC, `http://www.hcrc.ed.ac.uk/publications/rp-79.ps.gz`.

Corballis, M. C. 2007. Recursion, language, and starlings. *Cognitive Science 31,* 4, 697–704.

Córrea, L. M. 1995. An alternative assessment of children's comprehension of relative clauses. *Journal of Psycholinguistic Research 24*, 183–203.

Crain, S. and Nakayama, M. 1987. Structure dependence in grammar formation. *Language 63,* 3, 522–543.

Crain, S. and Pietroski, P. M. 2001. Nature, nurture, and universal grammar. *Linguistics and Philosophy 24*, 139–186.

Crain, S. and Thornton, R. 1998. Wanna contraction. In *Investigations in Universal Grammar: A Guide to Experiments in the Acquisition of Syntax and Semantics*, S. Crain and R. Thornton, Eds. MIT Press, Cambridge, MA, 177–186.

Crick, F. H. C. 1989. The recent excitement about neural networks. *Nature 337*, 129–132.

Dabrowska, E. 2000. From formula to schema: The acquisition of English questions. *Cognitive Linguistics 11*, 83–102.

Das, S., Giles, C. L., and Sun, G.-Z. 1992. Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In *Proceedings of the 14th Conference of the Cognitive Science Society.* Erlbaum, Bloomington, IN, 791–796.

Dell, G. S., Chang, F., and Griffin, Z. M. 1999. Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science 23,* 4, 517–542.

Demetras, M. J., Post, K. N., and Snow, C. E. 1986. Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language 13*, 275–292.

Diessel, H. 2004. *The Acquisition of Complex Sentences.* Cambridge Studies in Linguistics 105. Cambridge University Press.

——— 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology 25*, 108–127.

Diessel, H. and Tomasello, M. 2000. The development of relative clauses in spontaneous child speech. *Cognitive Linguistics 11,* 1/2, 131–151.

——— 2005. A new look at the acquisition of relative clauses. *Language 81,* 4, 882–906.

Dowty, D. 1991. Thematic proto-rules and argument selection. *Language 67*, 547–619.

Eimas, P. 1999. Do infants learn grammar with algebra or statistics? *Science 284*, 436.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science 14,* 2, 179–211.

——— 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning 7*, 195–225.

——— 1993. Learning and development in neural networks: The importance of starting small. *Cognition 48,* 1, 71–99.

Ferreira, F. 1994. Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language 33*, 715–736.

——— 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology 47*, 164–203.

Ferreira, F., Apel, J., and Henderson, J. M. 2008. Taking a new look at looking at nothing. *Trends in Cognitive Sciences.* To appear.

Ferreira, F. and Clifton, C. 1986. The independence of syntactic processing. *Journal of Memory and Language 25*, 348–368.

Ferreira, V. S. and Dell, G. S. 2000. The effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology 40*, 296–340.

Fillmore, C. J. 1968. The case for case. In *Universals in Linguistic Theory*, E. Bach and R. T. Harms, Eds. Holt, Rinehart and Winston, New York, 1–88.

Fisher, C., Gleitman, H., and Gleitman, L. R. 1991. On the semantic content of subcategorization frames. *Cognitive Psychology 23*, 331–392.

Fitch, W. T. and Hauser, M. D. 2004. Computational constraints on syntactic processing in a nonhuman primate. *Science 303,* 5656, 377–380.

Fitch, W. T., Hauser, M. D., and Chomsky, N. 2005. The evolution of the language faculty: Clarifications and implications. *Cognition 97,* 2, 179–210.

Fitz, H. 2007. Church's thesis and physical computation. In *Church's Thesis after 70 Years*, A. Olszewski and J. Wolenski, Eds. Ontos (Mathematical Logic), Berlin, 175–219.

Fitz, H. and Chang, F. 2008. The role of the input in a connectionist model of the accessibility hierarchy in development. In *Proceedings of the 32nd Annual Boston University Conference on Language Development*, H. Chan, H. Jacob, and E. Kapia, Eds. Vol. 1. Cascadilla, Somerville, MA, 120–131.

——— 2009. Syntactic generalization in a connectionist model of complex sentence production. In *Connectionist Models of Behavior and Cognition II. Proceedings of* *the 11th Neural Computation and Psychology Workshop, Oxford University, July 16-18 2008*, J. Mayor, N. Ruh, and K. Plunkett, Eds. Progress in Neural Processing, vol. 18. World Scientific Press, 289–300.

Fodor, J. and Garrett, M. 1967. Some syntactic determinants of sentential complexity. *Perception and Psychophysics 2*, 289–296.

Fodor, J. and McLaughlin, B. P. 1990. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition 35*, 183–204.

Fodor, J. A. and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition 28*, 3–71.

Forcada, M. L. and Carrasco, R. C. 1995. Learning the initial state of a second-order recurrent neural network during regular-language inference. *Neural Computation 7,* 5, 923–930.

Ford, M. and Holmes, V. M. 1978. Planning units and syntax in sentence production. *Cognition 6*, 35–53.

Foss, D. J. and Cairns, H. S. 1970. Some effects of memory limitation upon sentence comprehension and recall. *Journal of Verbal Learning and Verbal Behavior 9,* 5, 541–547.

Frank, R., Mathis, D., Gussine, E., Stowe, J., and Vindiola, M. 2006. Question formation, neural networks and the poverty of the stimulus. Poster presented at the *12th Annual Conference on Architectures and Mechanisms for Language Processing*, Nijmegen, `http://www.cog.jhu.edu/faculty/frank/papers/AMLAP-poster.pdf`.

Frazier, L. and Clifton, C. 1989. Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes 4,* 2, 93–126.

Frazier, L. and Rayner, K. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology 14*, 178–210.

Friedmann, N. and Novogrodsky, R. 2004. The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child Language 31,* 3, 661–681.

Friedrich, C. K. and Kotz, S. A. 2007. Event-related potential evidence of form and meaning coding during online speech recognition. *Journal of Cognitive Neuroscience 19,* 4, 594–604.

Funahashi, K.-I. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks 2*, 183–192.

Garrett, M. F. 1988. Processes in language production. In *Language: Psychological and Biological Aspects*, F. J. Newmeyer, Ed. Cambridge University Press, New York, 69–96.

Gennari, S. P. and MacDonald, M. C. 2008. Semantic indeterminacy in object relative clauses. *Journal of Memory and Language 58*, 161–187.

Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. 2006. Recursive syntactic pattern learning by songbirds. *Nature 440*, 1204–1207.

Gerken, L., Wilson, R., and Lewis, W. 2005. Infants can use distributional cues to form syntactic categories. *Journal of Child Language 32*, 249–268.

Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition 68*, 1–76.

Gigerenzer, G. 2000. *Adaptive Thinking: Rationality in the Real World.* Oxford University Press, New York.

Giles, C. L., Miller, C. B., Chen, D., Chen, H. H., Sun, G. Z., and Lee, Y. C. 1992. Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation. 4,* 3, 393–405.

Gold, M. E. 1967. Language identification in the limit. *Information and Control 10,* 5, 447–474.

Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure.* University of Chicago Press.

——— 1999. The emergence of argument structure semantics. In *The Emergence of Language*, B. MacWhinney, Ed. Erlbaum, Hillsdale, NJ, 197–212.

——— 2006. *Constructions at Work: The Nature of Generalization in Language.* Oxford University Press.

Goldberg, A. E. and Sethuraman, N. 2004. Learning argument structure generalizations. *Cognitive Linguistics 14*, 289–316.

Goldin-Meadow, S. and Mylander, C. 1998. Spontaneous sign systems created by deaf children in two cultures. *Nature 391*, 279–281.

Gómez, R. 2007. Statistical learning in infant language development. In *The Oxford Handbook of Psycholinguistics*, G. Gaskell, Ed. Oxford University Press.

Gómez, R. and Gerken, L. A. 1999. Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition 70*, 109–135.

Goodglass, H. and Kaplan, E. 1983. *The Assessment of Aphasia and Related Disorders.* Lea & Febiger, Philadelphia, PA.

Gordon, P. C., Hendrick, R., and Johnson, M. 2001. Effects of noun phrase type on sentence complexity. *Journal of Memory and Language 51,* 1, 97–114.

Gori, M., Maggini, M., Martinelli, E., and Soda, G. 1998. Inductive inference from noisy examples using the hybrid finite state filter. *IEEE Transactions on Neural Networks 9,* 3, 571–575.

Greenfield, P. M. 1991. Language, tools, and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and Brain Sciences 14,* 4, 531–551.

Griffin, Z. M. and Bock, K. 2000. What the eyes say about speaking. *Psychological Science 11*, 274–279.

Gropen, J., Pinker, S., Hollander, M., and Goldberg, R. 1991. Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition 41*, 153–195.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., and Wilson, R. 1989. The learnability and acquisition of the dative alternation in English. *Language 65*, 203–257.

Grossberg, S. 1973. Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics 52*, 213–257.

Grüning, A. 2007. Elman backpropagation as reinforcement for simple recurrent networks. *Neural Computation 19,* 3, 3108–3131.

Gupta, P. and Cohen, N. J. 2002. Theoretical and computational analysis of skill learning, repetition priming, and procedural memory. *Psychological Review 109,* 2, 401–448.

Hadley, R. F. 1994. Systematicity in connectionist language learning. *Mind and Language 9*, 247–272.

——— 2004. On the proper treatment of semantic systematicity. *Minds and Machines 14*, 145–172.

Hakes, D. T., Evans, J. S., and Brannon, L. L. 1976. Understanding sentences with relative clauses. *Memory and Cognition 4*, 283–290.

Hale, J. 2006. Uncertainty about the rest of the sentence. *Cognitive Science 30*, 643–672.

Hanna, J. E. and Tanenhaus, M. K. 2004. Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science 28*, 105–115.

Hauser, M. D., Chomsky, N., and Fitch, W. T. 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science 298*, 1569–1579.

Hawkins, J. A. 1994. *A Performance Theory of Order and Constituency.* Cambridge University Press.

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., and Rapoport, S. I. 1991. Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences - USA 88*, 1621–1625.

Heim, S., Opitz, B., Müller, K., and Friederici, A. D. 2003. Phonological processing during language production: fMRI evidence for a shared production-comprehension network. *Cognitive Brain Research 16*, 285–296.

Hinton, G. E. and McClelland, J. L. 1988. Learning representations by recirculation. In *Neural Information Processing Systems*, D. Z. Anderson, Ed. American Institute of Physics, New York, 358–366.

Hirsh-Pasek, K., Treiman, R., and Schneidermann, M. 1984. Brown and Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of Child Language 11*, 81–88.

Holmes, V. M. 1988. Hesitations and sentence planning. *Language and Cognitive Processes 3*, 323–361.

Hopper, P. J. and Thompson, S. A. 1980. Transitivity in grammar and discourse. *Language 56*, 251–299.

Horne, B. G. and Giles, C. L. 1995. An experimental comparison of recurrent neural networks. In *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds. MIT Press, Cambridge, MA, 697–704.

Hornik, K., Stinchcombe, M., and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks 2,* 5, 359–366.

Huang, M. S. 1983. A developmental study of children's comprehension of embedded sentences with and without semantic constraints. *Journal of Psychology 114*, 51–56.

Ishizuka, T. 2005. Processing relative clauses in Japanese. *UCLA Working Papers in Psycholinguistics* 2, 135–157.

Israel, M. 2002. Consistency and creativity in first language acquisition. In *Proceedings of the Berkeley Linguistics Society.* University of California, Berkeley, 123–135.

Jackendoff, R. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford University Press.

Jackendoff, R. and Pinker, S. 2005. The nature of the language faculty and its implications for evolution of language (reply to Fitch, Hauser, and Chomsky). *Cognition 9 7*, 211–225.

Jim, K., Giles, C. L., and Horne, B. G. 1996. An analysis of noise in recurrent neural networks: Convergence and generalization. *IEEE Transactions on Neural Networks 7,* 6, 1424–1438.

Johnson, E. K. and Jusczyk, P. W. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language 44*, 548–567.

Joshi, A. K. 1995. Tree-adjoining grammars: How much context sensitivity is required to provide reasonable structural descriptions? In *Natural Language Parsing*, D. Dowty, L. Karttunen, and A. Zwicky, Eds. Cambridge University Press, 206–250.

Just, M. A. and Carpenter, P. A. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review 99*, 122–149.

Kam, X., Stoyeshka, I., Tornyova, L., Fodor, J. D., and Sakas, W. G. 2005. Statistics vs. UG in language acquisition: Does a bigram analysis predict auxiliary inversion? In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition.* Association for Computational Linguistics, Ann Arbor, MI, 69–71.

——— 2007. Bigram-based learning and the richness of the stimulus for language acquisition. `http://web.gc.cuny.edu/dept/lingu/liba/papers/kam2007.pdf`.

Kamide, Y., Altmann, G. T. M., and Haywood, S. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language 49*, 133–156.

Kaschak, M. P. and Glenberg, A. M. 2000. Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language 43*, 508–529.

Keenan, E. L. and Comrie, B. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry 8,* 1, 63–99.

Keenan, E. L. and Hawkins, S. 1987. The psychological validity of the accessibility hierarchy. In *Universal Grammar: 15 Essays*, E. L. Keenan, Ed. Croon Helm, London.

Kidd, E. 2003. Relative clause comprehension revisited: Commentary on Eisenberg (2002). *Journal of Child Language 30*, 671–679.

Kidd, E. and Bavin, E. L. 2002. English-speaking children's comprehension of relative clauses: Evidence for general-cognitive and language-specific constraints on development. *Journal of Psycholinguistic Research 6,* 31, 599–617.

Kidd, E., Brandt, S., Lieven, E., and Tomasello, M. 2007. Object relatives made easy. *Language and Cognitive Processes 22*, 860–897.

Kilian, J. and Siegelmann, H. T. 1996. The dynamic universality of sigmoidal neural networks. *Information and Computation 128*, 48–56.

Kim, A. and Osterhout, L. 2005. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language 52*, 205–225.

King, J. and Just, M. A. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language 30,* 5, 580–602.

Knoeferle, P. and Crocker, M. W. 2008. The coordinated processing of scene and utterance: Evidence from eye tracking in depicted events. In *Advances in Cognitive Science*, N. Srinivasan, A. K. Gupta, and J. Pandey, Eds. Sage Publications.

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. 2005. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition 95,* 1, 95–127.

Kremer, S. C. 1995. On the computational power of Elman-style recurrent networks. *IEEE Transactions on Neural Networks 6,* 4, 1000–1004.

Krogh, A. and Hertz, J. A. 1992. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems 4*, J. E. Moody, S. J. Hanson, and R. P. Lippman, Eds. Morgan Kaufmann, San Mateo, CA, 950–957.

Kuno, S. 1976. Subject, theme, and the speaker's empathy—a reexamination of relativization phenomena. In *Subject and Topic*, C. Li, Ed. Academic Press, New York, 417–444.

Kwasny, S. C. and Kalman, B. L. 1995. Tail-recursive distributed representations and simple recurrent networks. *Connection Science 7,* 1, 61–80.

Labelle, M. 1996. The acquisition of relative clauses: Movement or no movement? *Language Acquisition 5*, 2, 65–82.

Lambalgen, M. van and Hamm, F. 2005. *The Proper Treatment of Events*. Explorations in Semantics. Blackwell.

Lambrecht, K. 1994. *Information Structure and Sentence Form. Topic, Focus and the Mental Representation of Discourse Referents*. Cambridge University Press.

Landau, B. and Jackendoff, R. 1993. 'What' and 'where' in spatial language and cognition. *Behavioral and Brain Sciences 16*, 217–238.

Larkin, W. and Burns, D. 1977. Sentence comprehension and memory for embedded structure. *Memory and Cognition 5*, 17–22.

Legate, J. A. and Yang, C. D. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review 19*, 151–162.

Levelt, W. J. M. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Levin, B. and Rappaport Hovav, M. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. MIT Press, Cambridge, MA.

Levine, R. D. 2001. The extraction riddle: Just what are we missing? *Journal of Linguistics 37*, 145–174.

Lewis, J. and Elman, J. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development*, B. Skarabela, S. Fish, and A. H. J. Do, Eds. Vol. 2. Cascadilla Press, Somerville, MA, 359–370.

Lewis, R. L. and Vasishth, D. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science 29*, 375–421.

Lidz, J., Waxman, S., and Freedman, J. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition 89*, B65–B73.

Love, T. 2007. The processing of non-canonically ordered constituents in long distance dependencies by pre-school children: A real-time investigation. *Journal of Psycholinguistic Research 36,* 3, 191–206.

Love, T. and Swinney, D. 1996. Coreference processing and levels of analysis in objectrelative constructions: Demonstration of antecedent reactivation with the cross-modal priming paradigm. *Journal of Psycholinguistic Research 20,* 1, 5–24.

Maass, W. and Orponen, P. 1998. On the effect of analog noise on discrete time analog computations. *Neural Computation 10*, 1071–1095.

MacDonald, M. C. and Christiansen, M. H. 2002. Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review 109,* 1, 35–54.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review 101,* 4 (10), 676–703.

MacNamara, J. 1972. Cognitive basis of language learning in infants. *Psychological Review 79,* 1, 1–13.

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

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

——— 2005. Item-based constructions and the logical problem. In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition.* Association for Computational Linguistics, Ann Arbor, MI, 53–68.

MacWhinney, B. and Pléh, C. 1988. The processing of restrictive relative clauses in Hungarian. *Cognition 29*, 95–141.

Mak, W. M., Vonk, W., and Schriefers, H. 2002. The influence of animacy on relative clause processing. *Journal of Memory and Language 47*, 50–68.

Manolios, P. and Fanelli, R. 1994. First-order recurrent neural networks and deterministic finite state automata. *Neural Computation 6,* 6, 1155–1173.

Marcus, G. 1993. Negative evidence in language acquisition. *Cognition 46,* 1, 53–85.

——— 1998. Rethinking eliminative connectionism. *Cognitive Psychology 37*, 243–282.

——— 1999a. Connectionism: With or without rules? *Trends in Cognitive Sciences 3*, 168–170.

——— 1999b. Do infants learn grammar with algebra or statistics? Response to Seidenberg & Elman, Negishi, and Eimas. *Science 284*, 436–437.

——— 1999c. Reply to Christiansen and Curtin. *Trends in Cognitive Sciences 3,* 8, 290–291.

Marcus, G. F., Vijayan, S., Bandi Rao, S., and Vishton, P. M. 1999. Rule learning by seven-month-old infants. *Science 283*, 77–80.

Marks, L. E. 1968. Scaling of grammaticalness of self-embedded English sentences. *Journal of Verbal Learning and Verbal Behavior 7,* 5, 965–967.

Marslen-Wilson, W., Levy, E., and Tyler, L. K. 1982. Producing interpretable discourse: The establishment and maintenance of reference. In *Speech, Place and Action. Studies in Deixis and Related Topics*, R. J. Jarvella and W. Klein, Eds. John Wiley, Chichester, 339–378.

Maskara, A. and Noetzel, A. 1992. Forcing simple recurrent neural networks to encode context. In *Proceedings of the 1992 Long Island Conference on Artificial Intelligence and Computer Graphics.* `http://www.funet.fi/pub/sci/neural/neuroprose/maskara.fsrn.ps.Z`.

Maslen, R. J. C., Theakston, A. L., Lieven, E., and Tomasello, M. 2004. A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research 47,* 6, 1319–1333.

Matsumoto, Y. 1999. Interaction of factors in construal: Japanese relative clauses. In *Grammatical Constructions: Their Form and Meaning*, M. Shibatani and S. A. Thompson, Eds. Oxford University Press, 103–124.

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

Mazzoni, P., Andersen, R. A., and Jordan, M. I. 1991. A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences 88*, 4433–4437.

McClelland, J. L. and Plaut, D. C. 1999. Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences 3,* 5, 166–168.

McRae, K., Ferretti, T. R., and Amyote, L. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes 12*, 137–176.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language 38*, 283–312.

Mecklinger, A. and Müller, N. 1996. Dissociations in the processing of 'what' and 'where' information in working memory: An event-related potential analysis. *Journal of Cognitive Neuroscience 8,* 5, 453–473.

Mecklinger, A. and Pfeifer, E. 1996. Event-related potentials reveal topographical and temporal distinct neuronal activation patterns for spatial and object working memory. *Cognitive Brain Research 4,* 3, 211–224.

Melnik, O., Levy, S., and Pollack, J. 2000. RAAM for infinite context-free languages. *Neural Networks 5*, 585–590.

Miikkulainen, R. 1990. A PDP architecture for processing sentences with relative clauses. In *Proceedings of the 13th Conference on Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, 201–206.

——— 1996. Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science 20*, 47–73.

——— 1997. Natural language processing with subsymbolic neural networks. In *Neural Network Perspectives on Cognition and Adaptive Robotics*. Institute of Physics Publishing, 120–139.

Miikkulainen, R. and Dyer, M. G. 1991. Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science 15,* 3, 343–399.

Miikkulainen, R. and Mayberry, M. R. 1999. Disambiguation and grammar as emergent soft constraints. In *Emergence of Language*, B. MacWhinney, Ed. Erlbaum, Hillsdale, NJ.

Miller, C. B. and Giles, C. L. 1993. Experimental comparison of the effect of order in recurrent neural networks. *International Journal of Pattern Recognition and Artificial Intelligence 7,* 4, 849–872.

Miller, G. A. 1962. Some psychological studies of grammar. *American Psychologist 17*,

748–762.

Miller, G. A. and Chomsky, N. 1963. Finitary models of language users. In *Handbook of Mathematical Psychology Vol. 2*, R. D. Lute, R. Bush, and E. Galanter, Eds. Wiley and Sons, New York, 419–91.

Miller, G. A. and Isard, S. 1964. Free recall of self-embedded English sentences. *Information and Control 7*, 292–303.

Minsky, M. and Papert, S. 1969. *Perceptrons.* MIT Press, Cambridge, MA.

Mintz, T., Newport, E. L., and Bever, T. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science 26*, 393–424.

Mintz, T. H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition 90*, 91–117.

Moerk, E. 1991. Positive evidence for negative evidence. *First Language 11,* 32, 219–251.

Mohri, M. and Nederhof, M.-J. 2001. Regular approximation of context-free grammars through transformation. In *Robustness in Language and Speech Technology*, J.-C. Junqua and G. van Noord, Eds. Kluwer, 153–163.

Moore, C. 1999. Dynamical recognizers: Real-time language recognition by analog computers. *Theoretical Computer Science 201*, 99–136.

Morgan, J. L., Bonamo, K. M., and Travis, L. L. 1995. Negative evidence on negative evidence. *Developmental Psychology 31,* 2, 180–197.

Morgan, J. L. and Travis, L. L. 1989. Limits on negative information in language input. *Journal of Child Language 16*, 531–552.

Nakano, Y., Felser, C., and Clahsen, H. 2002. Antecedent priming at trace positions in Japanese long-distance scrambling. *Journal of Psycholinguistic Research 31*, 531–571.

Ñeco, R. P. and Forcada, M. L. 1996. Beyond Mealy machines: Learning translators with recurrent neural networks. In *Proceedings of the World Conference on Neural Networks*. 408–411.

Negishi, M. 1999. Do infants learn grammar with algebra or statistics? *Science 284*, 435.

Nelson, K. E., Denninger, M. S., Bonvillian, J. D., Kaplan, B. J., and Baker, N. D. 1984. Maternal input adjustments and non-adjustments as related to children's linguistic advances and to language acquisition theories. In *The Development of Oral and Written Language in Social Contexts*, A. D. Pellegrini and T. D. Yawkey, Eds. Ablex, Norwood, NJ, 31–56.

Neto, J. P., Siegelmann, H. T., Costa, J. F., and Araujo, C. P. S. 1997. Turing universality of neural nets (revisited). In *Proceedings of the A Selection of Papers from the 6th International Workshop on Computer Aided Systems Theory*, F. Pichler and R. Moreno-Díaz, Eds. Lecture Notes in Computer Science, vol. 1333. Springer, London, 361–366.

Newmeyer, F. 2000. *Language Form and Language Function.* MIT Press, Cambridge, MA.

Newport, E. L. 1990. Maturational constraints on language learning. *Cognitive Science 14*, 11–28.

Odifreddi, P. 1989. *Classical Recursion Theory.* Studies in Logic, vol. 125. North-Holland, Amsterdam.

O'Grady, W., Nakamura, M., and Ito, Y. 2008. Want-to contraction in second language acquisition: An emergentist approach. *Lingua 118,* 4, 478–498.

Omlin, C. W. and Giles, C. L. 1996. Constructing deterministic finite-state automata in recurrent neural networks. *Journal of the ACM 43,* 6, 937–972.

Onnis, L. 2003. Statistical language learning. Ph.D. thesis, University of Warwick, England.

O'Reilly, R. C. 1996. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation 8,* 5, 895–938.

Osterhout, L. 1997. On the brain response to syntactic anomalies: Manipulation of word position and word class reveal individual differences. *Brain and Language 59*, 494–522.

Page, M. 2000. Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences 23*, 443–512.

Parsons, T. 1994. *Events in the Semantics of English: A Study in Subatomic Semantics.* MIT Press, Cambridge, MA.

Partee, B. 1965. Subject and object in modern English. In *Outstanding Dissertations in Linguistics Series*, J. Hankamer, Ed. Garland, New York.

Penner, S. 1987. Parental responses to grammatical and ungrammatical child utterances. *Child Development 58*, 376–384.

Perfors, A., Tenenbaum, J. B., and Regier, T. 2008. The learnability of abstract syntactic principles. Draft, `http://web.mit.edu/cocosci/Papers/perfors-tenenbaum-regier-submitted.pdf`.

Perruchet, P. and Rey, A. 2005. Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review 12*, 307–313.

Piattelli-Palmarini, M. 1980. *Language and Learning: The Debate between Jean Piaget and Noam Chomsky.* Harvard University Press, Cambridge, MA.

Pickering, M. J. and Barry, G. D. 1991. Sentence processing without empty categories. *Language and Cognitive Processes 6*, 229–259.

Pickering, M. J. and Branigan, H. P. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language 39*, 633–651.

Pickering, M. J. and Garrod, S. 2007. Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences 11,* 3, 105–110.

Pinker, S. 1984. *Language Learnability and Language Development: The Acquisition of Argument Structure.* Harvard University Press, Cambridge, MA.

——— 1989. *Learnability and Cognition: The Acquisition of Argument Structure.* MIT Press, Cambridge, MA.

——— 1990. Language acquisition. In *Language: An Invitation to Cognitive Science*, D. N. Osherson and H. Lasnik, Eds. Vol. 1. MIT Press, Cambridge, MA, 199–241.

Pinker, S. and Jackendoff, R. 2005. The faculty of language: What's special about it? *Cognition 95,* 2, 201–236.

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

Pollack, J. B. 1987. On connectionist models of natural language processing. Ph.D. thesis, University of Illinois, Urbana.

——— 1988. Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society.* Erlbaum, Hillsdale, NJ, 33–39.

——— 1990. Recursive distributed representations. *Artificial Intelligence 46*, 77–105.

——— 1991. The induction of dynamical recognizers. *Machine Learning 7*, 227–252.

Pullum, G. K. 1996. Learnability, hyperlearning, and the poverty of the stimulus. In *Proceedings of the 22nd Annual Meeting: General Session and Parasession on the Role of Learnability in Grammatical Theory*, J. Johnson, M. L. Juge, and J. L. Moxley, Eds. Berkeley Linguistics Society, 498–513.

Pullum, G. K. and Rogers, J. 2006. Animal pattern learning experiments: Some mathematical background. Unpublished draft, `http://ling.ed.ac.uk/~gpullum/index.html`.

Pullum, G. K. and Scholz, B. C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review 19,* 1, 9–50.

——— 2008. Recursion and the infinitude claim. Unpublished draft, `http://ling.ed.ac.uk/~gpullum/index.html`.

Pulvermüller, F. 1995. Agrammatism: Behavioral description and neurobiological explanation. *Journal of Cognitive Neuroscience 7,* 2, 165–181.

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

——— 2007a. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language 57*, 1–23.

——— 2007b. Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology 60*, 161–170.

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

Regier, T. and Gahl, S. 2004. Learning the unlearnable: The role of missing evidence. *Cognition 93*, 147–155.

Roberts, L., Marinis, T., Felser, C., and Clahsen, H. 2007. Antecedent priming at trace positions in children's sentence processing. *Journal of Psycholinguistic Research 36*, 175–188.

Rodriguez, P. 2001. Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation 13*, 2093–2118.

Rodriguez, P., Wiles, J., and Elman, J. L. 1999. A recurrent neural network that learns to count. *Connection Science 11,* 1, 5–40.

Rohde, D. 1999. Lens: The light, efficient network simulator. Tech. Rep. CMUCS-99-164, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

——— 2002. A connectionist model of sentence comprehension and production. Ph.D. thesis, Carnegie Mellon University, Pittsburgh.

Rohde, D. L. T. and Plaut, D. C. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition 72*, 67–109.

Rojas, R. 1996. *Neural Networks. A Systematic Approach.* Springer.

Rowland, C. F. and Pine, J. M. 2000. Subject-auxiliary inversion errors and wh-question acquisition: 'what children do know?'. *Journal of Child Language 27*, 157–181.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds. MIT Press, Cambridge, MA, 318–362.

Saffran, J. R. 2001. Words in a sea of sounds: The output of statistical learning. *Cognition 81*, 149–169.

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

Sag, I. A. 1997. English relative clause constructions. *Journal of Linguistics 33*, 431–483.

Sag, I. A. and Fodor, J. D. 1994. Extraction without traces. In *Proceedings of the Thirteenth Annual Meeting of the West Coast Conference on Formal Linguistics.* CSLI Publications, Stanford, 365–384.

Sampson, G. 1989. Language acquisition: Growth or learning? *Philosophical Papers 18*, 203–240.

Sanfeliu, A. and Alquézar, R. 1994. Active grammatical inference: A new learning methodology. In *Shape and Structure in Pattern Recognition*, D. Dori and A. Bruckstein, Eds. World Scientific Press, Singapore, 191–200.

Scheepers, C. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition 89*, 179–205.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition 71*, 109–148.

Seidenberg, M. S. and Elman, J. L. 1999a. Do infants learn grammar with algebra or statistics? *Science 284*, 434–435.

——— 1999b. Networks are not 'hidden rules'. *Trends in Cognitive Sciences 3,* 8, 288–289.

Servan-Schreiber, D., Cleeremans, A., and McClelland, J. L. 1991. Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning 7,* 2-3, 161–193.

Sheldon, A. 1974. The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior 13*, 272–281.

Siegelmann, H. T. 1999. *Neural Networks and Analog Computation: Beyond the Turing Limit.* Birkhäuser.

Siegelmann, H. T. and Sontag, E. D. 1991. Turing computability with neural nets. *Applied Mathematics Letters 4,* 6, 77–80.

Slobin, D. I. 1973. Cognitive prerequisites for the development of grammar. In *Studies of Child Language Development*, C. A. Ferguson and D. I. Slobin, Eds. Holt, Rinehart and Winston, New York, 175–208.

Smith, C. S. 1997. *The Parameter of Aspect*, 2nd ed. Kluwer, Norwell, MA.

Smith, E. E., Jonides, J., Koeppe, R. A., Awh, E., Schumacher, E. H., and Minoshima, S. 1995. Spatial versus object working memory: Pet investigations. *Journal of Cognitive Neuroscience 7*, 337–356.

Smith, M. 1974. Relative clause formation between 29-36 months: A preliminary report. *Papers and Reports on Child Language Development 8*, 104–110.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences 11*, 1–74.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., and Sedivy, J. C. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology 45*, 447–481.

Spivey-Knowlton, M. and Sedivy, J. C. 1995. Resolving attachment ambiguities with multiple constraints. *Cognition 55,* 3, 227–267.

Spivey-Knowlton, M. J. and Tanenhaus, M. K. 1998. Syntactic ambiguity resolution in discourse: Modelling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition 24*, 1521–1543.

St. John, M. F. and McClelland, J. L. 1992. Parallel constraint satisfaction as a comprehension mechanism. In *Connectionist Approaches to Natural Language Processing*, R. G. Reilly and N. E. Sharkey, Eds. Erlbaum, Hillsdale, NJ, 97–136.

Stabler, E. 1983. How are grammars represented? *Behavioral and Brain Sciences 6*,

391–402.

——— 2004. Varieties of crossing dependencies. *Cognitive Science 28,* 5, 699–720.

STEEDMAN, M. 1999. Connectionist sentence processing in perspective. *Cognitive Science 23,* 4, 615–634.

——— 2002. Plans, affordances and combinatory grammar. *Linguistics and Philosophy 25,* 5-6, 725–753.

STEIJVERS, M. AND GRÜNWALD, P. D. G. 1996. A recurrent network that performs a context-sensitive prediction task. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Erlbaum, Mahwah, NJ, 335–339.

STOLZ, W. S. 1967. A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior 6,* 6, 867–873.

STROMSWOLD, K., CAPLAN, D., ALPERT, N., AND RAUCH, S. 1996. Localization of syntactic comprehension by positron emission tomography. *Brain and Language 52,* 3, 452–473.

SUN, G.-Z., CHEN, H. H., GILES, C. L., LEE, Y. C., AND CHEN, D. 1990. Connectionist pushdown automata that learn context-free grammars. In *Proceedings of the International Joint Conference on Neural Networks*, M. Caudill, Ed. Vol. I. Erlbaum, Hillsdale, NJ, 577–580.

SWINNEY, D., FORD, M., FRAUENFELDER, U., AND BRESNAN, J. 1988. On the temporal course of gap-filling and antecedent assignment during sentence comprehension. In *Language Structure and Processing*, B. Grosz, R. Kaplan, M. Macken, and I. Sag, Eds. CSLI, Stanford, CA.

TALMY, L. 2000. *Toward a Cognitive Semantics.* Vol. 1. MIT Press, Cambridge, MA.

TAVAKOLIAN, S. 1981. The conjoined-clause analysis of relative clauses. In *Language Acquisition and Linguistic Theory*, S. Tavakolian, Ed. MIT Press, Cambridge, MA.

TAYLOR, J. R. 2002. *Cognitive Grammar.* Oxford University Press.

THOMAS, M. 2002. Development of the concept of "the poverty of the stimulus". *The Linguistic Review 19,* 1, 51–71.

THOMPSON, S. A. 1987. Subordination and narrative event structure. In *Coherence and Grounding in Discourse*, R. S. Tomlin, Ed. John Benjamins Publishing, 435–454.

TIŇO, P., HORNE, B. G., GILES, C. L., AND COLLINGWOOD, P. C. 1998. Finite state machines and recurrent neural networks - automata and dynamical systems approaches. In *Neural Networks and Pattern Recognition*, J. E. Dayhoff and O. Omidvar, Eds. Academic Press, 171–220.

TIŇO, P. AND SAJDA, J. 1995. Learning and extracting initial mealy automata with a modular neural network model. *Neural Computation 7,* 4, 822–844.

TOMASELLO, M. 1992. *First Verbs: A Case Study of Early Lexical Development.* Cambridge University Press.

——— 1999. *The Cultural Origins of Human Cognition.* Harvard University Press.

——— 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

Tomlin, R. S. 1985. Foreground-background information and the syntax of subordination. *Text 5*, 85–122.

Tonkes, B., Blair, A., and Wiles, J. 1998. Inductive bias in context-free language learning. In *Proceedings of the Ninth Australian Conference on Neural Networks.* 52–56.

Townsend, D. J. and Bever, T. G. 1978. Interclausal relations and clausal processing. *Journal of Verbal Learning and Verbal Behavior 17*, 509–521.

Traxler, M., Morris, R., and Seely, R. 2002. Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language 47*, 69–90.

Trueswell, J. C. and Tanenhaus, M. K. 1994. Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In *Perspectives on Sentence Processing*, C. Clifton, K. Rayner, and L. Frazier, Eds. Erlbaum, Hillsdale, NJ, Chapter 7.

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S. M. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language 33*, 285–318.

Ullman, M. T. 2001. A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews 2*, 717–727.

Ungerleider, L. G. and Mishkin, M. 1982. Two cortical visual systems. In *Analysis of Visual Behavior*, D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds. MIT Press, Cambridge, MA, 549–586.

Van Valin, R. 1998. The acquisition of *wh*-questions and the mechanisms of language acquisition. In *The New Psychology of Language. Cognitive and Functional Approaches to Language Structure*, M. Tomasello, Ed. Vol. 1. Erlbaum, Mahwah, NJ.

Verhagen, A. 2008. Syntax, recursion, productivity—a usage-based perspective on the evolution of grammar. In *Evidence and Counter-Evidence: Essays in Honour of Frederik Kortlandt Vol. 2*, A. Lubotsky, J. Schaeken, and J. Wiedenhof, Eds. Rodopi, Amsterdam.

Villiers, J. G. de, Flusberg, H., Hakuta, K., and Cohen, M. 1979. Children's comprehension of relative clauses. *Journal of Psycholinguistic Research 8,* 5, 499–518.

Wanner, E. and Maratsos, M. 1978. An ATN approach to comprehension. In *Linguistic Theory and Psychological Reality*, M. Halle, J. Bresnan, and J. Miller, Eds. Cambridge University Press, 119–161.

Warren, T. and Gibson, E. 2002. The influence of referential processing on sentence complexity. *Cognition 85*, 79–112.

Watkins, K. E., Strafella, A. P., and Paus, T. 2003. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia 41*, 989–994.

Watrous, R. L. and Kuhn, G. M. 1992. Induction of finite-state languages using second-order recurrent networks. *Neural Computation 4,* 3, 406–414.

Weckerly, J. and Elman, J. L. 1992. A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ, 414–419.

Wells, J., Christiansen, M. H., MacDonald, M., and Race, D. 2008. Experience and sentence comprehension: Implicit learning, working memory, and individual differences. *Cognitive Psychology*. Under revision.

Whitehurst, G. J. and Valdez-Menchaca, M. C. 1988. What is the role of reinforcement in early language acquisition? *Child Development 59,* 2, 430–440.

Wiles, J., Blair, A. D., and Bodén, M. 2001. Representation beyond finite states: Alternatives to push-down automata. In *A Field Guide to Dynamical Recurrent Networks*, J. F. Kolen and S. C. Kremer, Eds. IEEE Press, 129–142.

Wiles, J. and Elman, J. L. 1995. Learning to count without a counter: A case study of dynamics and activation landscapes in recurrent networks. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. MIT Press, Cambridge, MA, 482–487.

Xie, X. and Seung, H. S. 2003. Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation 15*, 441–454.

Zipser, D. and Andersen, R. A. 1988. A back propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature 331,* 6158, 679–684.

Samenvatting

Kinderen leren hun moedertaal spontaan en zonder enige moeite door interactie met hun omgeving; het is onnodig om hen expliciet de taal te leren. De taalervaring waaruit kinderen moeten leren is echter in hoge mate onbepaald en beargumenteerbaar ontoereikend met betrekking tot het leerdoel. Desondanks zullen de meeste zich normaal ontwikkelende kinderen hun moedertaal snel en met groot gemak leren spreken.

Veel taalverwervingstheorieën zoeken de verklaring hiervoor in aangeboren beperkingen van het grammaticale 'zoekgebied', of zien zelfs een biologische taalspecifieke predispositie. Gebruiksgebaseerde theorieën van taal daarentegen leggen meer nadruk op de rol die ervaring speelt en op domein-algemeene leermechanismen dan op aangeboren taalspecifieke kennis. Echter, talen zijn lexicaal onbeperkt en structureel eindeloos te combineren, dus hun uitdrukkingskracht is niet door middel van ervaring volledig te vangen. Gebruiksgebaseerde theorieën zullen daarom moeten verklaren hoe kinderen in staat zijn om de eigenschappen van hun taalinformatie te generaliseren naar een volwassen grammatica.

In deze dissertatie presenteer ik een expliciet computationeel mechanisme, waarmee de gebruiksgebaseerde theorieën van taal getest en geëvalueerd kunnen worden. De nadruk van mijn werk ligt op het gebied van complexe syntax en het menselijk vermogen om zinnen te vormen die meer dan één bewering uitdrukken door middel van bijzinsconstructies. Deze capaciteit voor recursie is een essentieel kenmerk van een volwassen grammatica en, zoals sommigen hebben beargumenteerd, van menselijke taal zelf.

De dissertatie is als volgt georganiseerd. Na een introductie geef ik in het tweede hoofdstuk een overzicht van resultaten, die de wiskundige eigenschappen van neurale netwerken karakteriseren en herzie ik eerder onderzoek in het modelleren van de verwerving van complexe syntax met zulke netwerken. Het hoofdstuk schetst daarmee het conceptuele landschap waarin het huidige werk zich bevindt.

In een derde hoofdstuk beargumenteer ik dat de constructie en het gebruik van betekenis essentieel is, in zowel kindertaalverwerving als volwassen taalverwerking, en dat neurale netwerkmodellen deze dimensie van menselijk taalgedrag moeten opnemen.

Ik introduceer het Dual-path model van zinsproductie en syntactische ontwikkeling. Het model is in staat om semantiek te representeren en het leert van invoer van zinnen gepaard aan hun betekenis (cf. Chang et al. 2006). Ik leg de architectuur van het model uit, geef de motivatie voor basisaannamen in het ontwerp, en bespreek bestaand onderzoek dat is uitgevoerd met het model.

Een vierde hoofdstuk beschrijft en vergelijkt enkele uitbreidingen van de basisarchitectuur die gericht zijn op de verwerking van uitingen met meerdere bijzinnen. Deze uitbreidingen worden geëvalueerd op basis van computationele *desiderata*, zoals bepaalde leer- en generaliseringsprestaties en de spaarzaamheid van semantische representaties. Een optimale oplossing voor het coderen van betekenis van complexe zinnen met betrekkelijke bijzinnen is vastgesteld. Dit vormt de basis voor alle verdere simulaties.

Hoofdstuk vijf analyseert de leerdynamiek van het model in meer detail. Eerst wordt het gedrag van het model voor verschillende types betrekkelijke bijzinnen bestudeerd. Syntactische varianten (zoals actief/passief) blijken bijzonder moeilijk te zijn, omdat ze de relatie tussen vorm en betekenis, die het model moet leren, ingewikkelder maken. In het tweede deel van het hoofdstuk kijk ik naar de interne representaties die het model ontwikkeld heeft tijdens leren. Ik beweer dat het model de argumentstructuur verwerft van de constructievormen in de invoertaal, en dat het de hiërarchische structuren van verschillende complexe uitingen representeert.

De kern van dit proefschrift is te vinden in de hoofdstukken zes tot en met acht. In hoofdstuk zes wordt het generaliseringsvermogen van het Dual-path model getoetst in diverse taken. Ik laat zien dat de syntactische representaties voldoende transparant zijn om structurele generalisatie naar nieuwe complexe uitingen mogelijk te maken. Semantische gelijkenissen tussen nieuwe en reeds bekende zinstypen spelen een cruciale rol in deze taak. Het Dual-path model heeft ook het vermogen om bekende woorden in nieuwe argumentposities in nieuwe constructies te kunnen generaliseren. Dit wordt 'sterke semantische systematiciteit' genoemd. Daarnaast stel ik leeromstandigheden vast waaronder het model recursieve productiviteit toont. Ik beargumenteer dat het gedrag van het model te vergelijken is met menselijk gedrag, in zoverre de nauwkeurigheid van productie vermindert met de diepte van de ingebedde bijzinnen, en rechts-ingebedde structuren sneller worden geleerd dan centraal-ingebedde structuren.

In hoofdstuk zeven bestudeer ik het leren van complexe ja/nee-vragen in de afwezigheid van voorbeelden in de input. Ik laat zien dat het Dual-path model de syntax van zulke vragen kan verwerven uit soortgelijke en eenvoudigere structuren, waarvan de aanwezigheid is aangetoond in de taalomgeving van kinderen. De fouten van het model zijn vergelijkbaar met de fouten die kinderen maken, en ik stel voor dat er geen taalspecifieke aanleg in kinderen moet worden verondersteld in het leren van complexe ja-nee vragen. Deze resultaten zijn relevant voor het *poverty of the stimulus* debat, omdat het model geen traditioneele universele grammatica implementeert.

Engelse bijzinsconstructies geven aanleiding tot vergelijkbare prestatierangschikkingen in volwassen taalverwerking en kindertaalverwerving. Dit patroon komt overeen met het typologische universeel die de '*noun phrase accessibility hierarchy*' wordt genoemd. In hoofdstuk acht stel ik een inputgebaseerde verklaring voor van deze observatie. Het Dual-path model laat deze rangschikking zien in de syntactische ontwikkeling wanneer het leert van plausibele inputdistributies. Het is echter mogelijk deze rangschikking te manipuleren en volledig te elimineren door de eigenschappen van de input te variëren. Ik beweer dat patronen van interferentie en vereenvoudiging tussen inputstructuren de hiërarchie kunnen verklaren wanneer alle structuren simultaan worden geleerd en gerepresenteerd over een enkele verzameling van neurale verbindingen.

Tot besluit trek ik conclusies uit mijn werk, signaleer een aantal onbeantwoorde vragen, en geef een korte vooruitblik op mogelijke onderzoeksuitbreidingen.

Titles in the ILLC Dissertation Series:

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander van den Bosch**
Rationality in Discovery - a study of Logic, Cognition, Computation and Neuropharmacology

ILLC DS-2001-03: **Erik de Haas**
Logics For OO Information Systems: a Semantic Study of Object Orientation from a Categorial Substructural Perspective

ILLC DS-2001-04: **Rosalie Iemhoff**
Provability Logic and Admissible Rules

ILLC DS-2001-05: **Eva Hoogland**
Definability and Interpolation: Model-theoretic investigations

ILLC DS-2001-06: **Ronald de Wolf**
Quantum Computing and Communication Complexity

ILLC DS-2001-07: **Katsumi Sasaki**
Logics and Provability

ILLC DS-2001-08: **Allard Tamminga**
Belief Dynamics. (Epistemo)logical Investigations

ILLC DS-2001-09: **Gwen Kerdiles**
Saying It with Pictures: a Logical Landscape of Conceptual Graphs

ILLC DS-2001-10: **Marc Pauly**
Logic for Social Software

ILLC DS-2002-01: **Nikos Massios**
Decision-Theoretic Robotic Surveillance

ILLC DS-2002-02: **Marco Aiello**
Spatial Reasoning: Theory and Practice

ILLC DS-2002-03: **Yuri Engelhardt**
The Language of Graphics

ILLC DS-2002-04: **Willem K. van Dam**
On Quantum Computation Theory

ILLC DS-2002-05: **Rosella Gennari**
Mapping Inferences: Constraint Propagation and Diamond Satisfaction

ILLC DS-2002-06: **Ivar Vermeulen**
A Logical Approach to Competition in Industries

ILLC DS-2003-01: **Barteld Kooi**
Knowledge, chance, and change

ILLC DS-2003-02: **Elisabeth Catherine Brouwer**
Imagining Metaphors: Cognitive Representation in Interpretation and Understanding

ILLC DS-2003-03: **Juan Heguiabehere**
Building Logic Toolboxes

ILLC DS-2003-04: **Christof Monz**
From Document Retrieval to Question Answering

ILLC DS-2004-01: **Hein Philipp Röhrig**
Quantum Query Complexity and Distributed Computing

ILLC DS-2004-02: **Sebastian Brand**
Rule-based Constraint Propagation: Theory and Applications

ILLC DS-2004-03: **Boudewijn de Bruin**
Explaining Games. On the Logic of Game Theoretic Explanations

ILLC DS-2005-01: **Balder D. ten Cate**
Model theory for extended modal languages

ILLC DS-2005-02: **Willem-Jan van Hoeve**
Operations Research Techniques in Constraint Programming

ILLC DS-2005-03: **Rosja Mastop**
What can you do? Imperative mood in Semantic Theory

ILLC DS-2005-04: **Anna Pilatova**
A User's Guide to Proper names: Their Pragmatics and Semanics

ILLC DS-2005-05: **Sieuwert van Otterloo**
A Strategic Analysis of Multi-agent Protocols

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian T. Semmes**
A Game for the Borel Functions