Syntactic Learning from Ambiguous Evidence: Errors and End-States

by

Isaac Gould

B.A. in Linguistics, The University of Toronto, 2009
M.A. in Linguistics, The University of Toronto, 2010

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Linguistics

at the

Massachusetts Institute of Technology

September 2015

Signature of Author: …….
Department of Linguistics and Philosophy
June 11, 2015

Certified by: ……..
Adam Albright
Associate Professor of Linguistics
Thesis Supervisor

Accepted by: …….
David Pesetsky
Ferrari P. Ward Professor of Linguistics
Head, Department of Linguistics and Philosophy

Syntactic Learning from Ambiguous Evidence: Errors and End-States

by

Isaac Gould

Submitted to the Department of Linguistics and Philosophy
on June 11, 2015 in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics

Abstract

In this thesis I explore the role of ambiguous evidence in first language acquisition by using a probabilistic learner for setting syntactic parameters. As ambiguous evidence is input to the learner that is compatible with multiple grammars or hypotheses, it poses learnability and acquisition challenges because it underdetermines the correct analysis. However, a probabilistic learning model with competing hypotheses can address these challenges by learning from general tendencies regarding the shape of the input, thereby finding the most compatible set of hypotheses, or the grammar with the 'best fit' to the input. This enables the model to resolve the challenge of learning the grammar of a subset language: it can reach such a target end-state by learning from implicit negative evidence. Moreover, ambiguous evidence can provide insight into two phenomena characteristic of language acquisition: variability (both within speakers and across a population) and learning errors. Both phenomena can be accounted for under a model that is attempting to learn a grammar of best fit.

Three case studies relating to word order and phrase structure are investigated with simulations of the model. First I show how the model can account for embedded clause verb placement errors in child Swiss German by learning from ambiguous input. I then show how learning from ambiguous input allows the model to account for grammatical variability across speakers with regard to verb movement in Korean. Finally, I show that the model is successfully able to learn the grammar of a subset language with the example of zero-derived causatives in English.

Thesis Supervisor: Adam Albright
Title: Associate Professor of Linguistics

Acknowledgments

The work in this thesis owes a considerable amount to a number of people. Above all, I would like to thank my committee members: Adam Albright, Michel DeGraff, and David Pesetsky.

From the very beginning, Adam helped me take some nebulous ideas and turn them into a research program. I have benefitted from his clarity regarding a diverse range of topics and his ability to patiently explain them, as well as his vision for the overall project and his interest in it throughout. I feel that I have learned a great deal from him.

I am thankful to have had the opportunity to work with David and Michel, who provided both valuable insight, as well as ongoing encouragement and words of support. My discussions with them have enriched the content and scope of this work and, I hope, have led to a document that is easier to read than it once was.

Special thanks also go to Nicholas Revett Rolle, Michelle Alison Fullwood, and Michael Yoshitaka Erlewine. Nik, perhaps unbeknownst to him, helped guide my nascent interest in acquisition and learnability toward the research reported here. To think what a discussion while walking through Golden Gate Park would lead to! Michelle was clutch in matters Church-y and was ever so helpful in patiently explaining to me things about modeling. mitcho, indefatigable throughout the course of long conversations, was a source of much inspiration, some of which helped to transform a huge chunk of this thesis.

At the risk of omission, I would like to thank the following people for having suggested ideas for research topics, shared their ideas to me about this research, talked to me about how to write it up, or for having simply listened to me talk about it: Samer Al Khatib, Artur Bartnikart, Isa Kerem Bayirli, Ailis Cournane, Edward Flemming, Samuel Freedman, Martin Hackl, Chung-hye Han, Aron Hirsch, Sabine Iatridou, Kwang-sup Kim, Hadas Kotek, Takashi Morita, Junya Nomura, Timothy O'Donnell, Myung-Kwan Park, Michelle Sheehan, Milena Sisovics, Mark Steedman, Ayaka Sugawara, Coppe van Urk, Kenneth Wexler, Sidney Winward, Patrick Wong, Suyeon Yun, and Hedde Zeijlstra.

And finally, I would like to express my love and affection to all those I have met and known over the last five years.

---

[*] This work has been supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Parts of this work have also benefitted from presentations at LingBaW 2 (Lublin, November 2014), PACLIC 28 (Phuket, December 2014), and WCCFL 33 (Vancouver, March 2015), as well as from various anonymous reviewers.

Contents

Chapter 1

Introduction

1. Introduction

How does a child learn properties of a language's syntax when the evidence in the surrounding linguistic environment that crucially bears on these syntactic properties is ambiguous? That is, how does the child learn in situations when the evidence is compatible with multiple hypotheses? This is the challenge presented by ambiguous evidence – the evidence available to the child underdetermines the correct structural analysis. The focus of this work is to highlight the role that ambiguous evidence plays in the process of language acquisition. Not only is ambiguous evidence highly prevalent in the child's dataset, but it can be highly informative as well. Moreover, learning from ambiguous evidence has implications for our understanding of two empirical phenomena that characterize the learning process: learner errors and variability in learners' grammars.

I propose a probabilistic learning model that is able to learn from the statistical tendencies of ambiguous data. A brief description of the model is given below, but I refer the reader to Chapter 2 for a more detailed description of the model, as well as a comparison with other prominent learning models. Notable points of difference with other models include the fact that the model here does not filter ambiguous evidence from the input to the learner, and that the model here uses ambiguous evidence in a principled way to move toward learning particular grammars. A strength of the model here is that it is able to learn from implicit negative evidence. Learning from implicit negative evidence is when the learner is sensitive to the absence of certain kinds of data in the input and infers that these data are not attested because they are not possible in the target language. Further, I present a novel insight about how a learner's hypotheses interact with each other that sheds light on learner errors and grammatical variability (see Section 3 and Chapters 3 and 4). The model is illustrated with a series of proof-of-concept case studies from three different languages in Chapters 3-5. Short previews of these case studies are presented in this chapter.

More precisely, the model I propose is a generative learner. That is, the model generates sentences (string-meaning pairs). Suppose the model is trying to learn a particular language *L* and is exposed to utterances from *L*. The grammar that has the greatest likelihood of generating the sentences that are actually found in *L* is the grammar that the model will end up learning. The objective of a generative learner is to maximize the likelihood of generating a particular data set (the utterances it has been exposed to) given a particular choice in grammar.

A brief description of how the model learns is as follows. I assume that a child's hypotheses about syntax can be represented by, and are constrained by a set of choices concerning parameter values (Chomsky 1981, 1986) and the application of phrase structure rules. Together, these choices allow the model to generate sentences. A parameter may have any number of values to which it can be set, although in all the cases I consider, the parameters will be binary. A key aspect of the probabilistic learner is the way that parameters are set. All of a given parameter's values have probabilistic weights assigned to them. Each weight represents the learner's expectation that that value is the target setting in the adult grammar. These expectations are based on the linguistic evidence the child learns from. As the null hypothesis I assume that a parameter's weights are equal at the beginning of the learning process. Similarly the application of phrase structure rules is associated with its own set of probabilistic weights that correspond to expectations. These weights determine which choices the learner makes to generate sentences. Throughout the learning process the learner's expectations (and thus the weights) are gradually adjusted (or reinforced) so as to better fit the input, and thus more and more closely approach the adult grammar. The greater the likelihood of generating the sentences found in the adult grammar, the closer the learner is to having acquired the adult grammar. Thus the learning process can be characterized as a competition among grammars for which one is likely to be the best fit to the evidence the learner has been exposed to. At a given point in time, each grammar has a different likelihood of fitting the evidence, and this is determined by the weights of all the choices (i.e. parameters and phrase structure rules) that comprise that grammar. Often there are multiple grammars that have high likelihoods. To be precise, the learner's hypothesis space is populated by parameters whose values have weights (not entire grammars themselves; grammars do not have their own weights), and the values of each parameter are engaged in a parameter-internal competition to push the learner toward a grammar of best fit to the input. The idea of grammar competition can thus be understood as emerging from the full network of these smaller parameter-internal competitions. In sum, the notions of grammar competition and a grammar of best fit will play a crucial role in learning from ambiguous evidence, as well as in modeling learner errors and variability.

In the remainder of this chapter, I first lay out an overview of some of the puzzles presented by ambiguous data, how they might be addressed by the probabilistic learning model proposed here, and how both model and data are related to learner errors and variability. This overview previews and summarizes the different kinds of case studies that are taken up in more detail in subsequent chapters.

2. The puzzle of ambiguous evidence

This section lays out a general presentation of the puzzle of ambiguous evidence and its various forms. I begin in Section 2.1 by discussing the question of how a learner might respond to ambiguous evidence. This is a crucial point upon which many learning models differ, and which plays a pivotal role in how the learner acquires a grammar. I then describe several scenarios in which ambiguous evidence can arise, the general case in Section 2.2 and a more specific case (the subset case) in Section 2.3. In this section I also lay out very generally how the model learns from ambiguous evidence. Discussing the general case in particular will set the stage in Section 3 for linking the learning model and ambiguous evidence to the acquisition phenomena of learner errors and variability.

2.1 Preliminary considerations

The main focus of this thesis is to address the question of what a learner does when encountering a datum of ambiguous evidence. If multiple grammars are compatible with the input, what does the learner do? A variety of approaches have been proposed in the literature. For example, the learner could give credit, or fully reinforce, all the compatible grammars (Boersma 1997). Or one could give partial credit (i.e. partially reinforce) all the compatible grammars (Magri 2013). Another approach would be to simply ignore the ambiguous evidence (Sakas and Fodor 2001; Pearl and Lidz 2009). Finally the learner could concentrate the credit on a single grammar or subset of grammars, in particular the subset of grammars that could be determined to be most compatible with the input (Yang 2002; see also Clark 1992; Clark and Roberts 1993). This last possibility is the sort of general approach I will pursue.[1]

What the learner does in response to ambiguous evidence is a crucial component of the learning process. Ultimately we want the child to learn the adult or target grammar (or something relatively close in the case of diachronic change). However, the target grammar is only one of potentially many grammars that are compatible with ambiguous input. If the learner were to reinforce a non-target grammar too much (or even at all), they might never be able to learn the target grammar, and they could end up learning some grammar that is wildly different from the target grammar (cf. Gold 1967; Clark 1989; Gibson and Wexler 1994; Sakas and Fodor 2012).

One prominent response to this challenge has been to have the learner rely exclusively on unambiguous evidence, or 'triggers', when setting parameters (e.g. Fodor 1998; Sakas and Fodor 2001). In general, the intended outcome of relying on unambiguous evidence is that the learner's grammar would consist of a monotonically

[1] An in-depth comparative discussion of all these approaches goes beyond the scope of this work. In this thesis, I will focus on comparing the probabilistic learner I propose with several well-articulated models of learning *syntactic* parameters. These alternative models in turn represent a variety of different approaches for what a learner does when faced with ambiguous evidence. Discussion of these models can be found in Chapter 2 and following the case studies in Chapters 3-5.

increasing set of parameters that are set correctly. In this way the learner can safely arrive at the target grammar. An approach that discards ambiguous evidence and does not learn from it can be called a deterministic learner. However, a deterministic learner raises several important acquisition questions. If learners are so attuned to unambiguous evidence and are simply learning a series of correct parameter settings, then why is it that children make errors that reflect non-target parameter values when learning their first language? Moreover, what would the learner do when there is never any (or vanishingly rare) unambiguous evidence for a particular parameter setting? And how could children learning the same language set a given parameter differently if they are always being guided by unambiguous evidence to a single grammar? At first glance, it is not obvious how a deterministic learner could address these questions, although I will consider and reject a possibility below.

My response to these questions is to adopt a learning model that attempts to find a grammar of best fit given both ambiguous and unambiguous evidence. The model does this in an online fashion and is constantly updating what this grammar of best fit is throughout the learning process. I will show that this kind of learning model provides insights into the answers to these questions. For example, before adopting the adult grammar, earlier in the learning process the grammar of best fit can be some other non-target grammar. Further, even when there is insufficient unambiguous evidence, a single grammar of best fit can emerge given the shape of the input. Crucially, it is learning from ambiguous evidence that results in a principled account of these learning questions.

It is worth highlighting the difference in approach between a deterministic learner (which learns only from unambiguous evidence), and my learning model, which learns from both ambiguous and unambiguous evidence. This difference crucially relates to the acquisition questions above: (a) how is that children make errors?; (b) what does the learner do when there is no (or vanishingly rare) unambiguous evidence?; and (c) how can children end up learning slightly different grammars as end-states? These questions form the core of this work. Looking more closely at a deterministic learning model can be instructive by throwing into sharp relief my answer to these questions: the necessity of learning from ambiguous evidence. This is because there is no formal role in the learning process for ambiguous evidence in a deterministic model, whereas it plays an integral role in mine.

As concerns language acquisition within a Principles and Parameters framework, then, a deterministic model presents essentially as stark a contrast as possible to my learning model. I do not deny that relying exclusively on unambiguous evidence is a sensible way to approach the challenge of language learnability (e.g. Gold 1967). It is therefore useful to investigate how a deterministic model might be shored up so as to address the acquisition questions I have raised. If there is some way of maintaining such an approach in the face of these questions, then perhaps ambiguous evidence need not play a role after all. In broad strokes below, I will now outline what seems to me the most

promising way of shoring up a deterministic approach to learning. This will involve the use of a universal set of default value for parameters. I will conclude that using defaults is insufficient in addressing these acquisition questions. More detailed discussion on defaults can be found in Chapter 6.

In an attempt to have a deterministic learner address these acquisition questions, let us consider augmenting such a learner with a universal set of defaults for all parameters. A default is a parameter value that the learner uses in the initial state. In a deterministic learner, a non-default value is adopted only if the learner hears sufficient unambiguous evidence for the non-default value. Along the lines of Sakas and Fodor (2012), I note that for the purpose of discussion here, the value of a default is not related to anything in particular such as markedness or derivational economy (according to which a movement parameter that is set to not have movement would be more economical than being set to have movement; cf. Clark and Roberts 1993). This is a more permissive view of what a default value might be: any of a parameter's values might be the value that all learners use as a default. This view of defaults gives us the greatest chance of finding a universal set of defaults that will allow a deterministic learner to account for language acquisition cross-linguistically. Accordingly, if we conclude that such a set of defaults (which can account for acquisition) cannot be found, this conclusion is a much stronger argument against a deterministic learner than a comparable conclusion would be with a more restrictive set of defaults. In this section, I discuss how even this permissive approach to defaults is inadequate in accounting for certain kinds of variability and errors we see across learners.

First, though, an immediate advantage of adopting defaults is the possibility of success when the learner will never encounter any unambiguous evidence for setting a particular parameter. Perhaps the most well known example of this involves what can be called the subset learning scenario (cf. Gold 1967; Wexler and Manzini 1987; and see Section 2.3 below). In this scenario, the input (and thus the target) is a subset language, all the evidence of which is also compatible with a superset language. For simplicity, suppose there is a single parameter P_1 whose different values, x and y, distinguish the grammars of the two languages: $P_1(x)$ gives us the subset language, and $P_1(y)$ the superset language. The learner never hears any unambiguous evidence for $P_1(x)$ or $P_1(y)$, but a default parameter value that has the learner start off with $P_1(x)$ will allow for the possibility of the learner converging on the grammar of the subset language, given that the learner never hears any unambiguous evidence for $P_1(y)$ (cf. Berwick 1986). Thus to ensure that the grammar of the subset language could always be learned if it is the target, a deterministic learner would need a universal default for the grammar of the subset language, i.e. $P_1(x)$. In contrast, the probabilistic learner I propose can converge on the grammar of the subset language by a process of inference. Given that the attested evidence is always compatible with the grammars of both the subset and superset languages, and given that evidence only compatible with the grammar of the superset

language is never attested, the model can learn from the ambiguous input that the grammar of the subset language is the better fit to the data.

The absence of unambiguous evidence goes beyond the subset scenario just described. In Chapter 4, I discuss how a child learning a verb-final language such as Korean is faced with a number of parameters related to word order for which it appears there is no (or vanishingly rare) unambiguous evidence. In this scenario, there are several parameters with multiple values that are all compatible with (essentially) all the input, and no grammar generates a language that can be considered a proper subset of any of the others. Given this insufficiency of unambiguous evidence, no parameter can ever (or virtually never) be set with a deterministic learner. However, a deterministic learner could have an initial state with a set of default values that form a grammar that is consistent with the input. To ensure that all learners start in such an initial state (given that no parameters can be set in this scenario, and thus it is not possible to move from a non-target state) the set of default values would again need to be universal. The widespread use of defaults is not necessary for the probabilistic learner I propose. This learner can capitalize on the fact that not all combinations of parameter values are compatible with the input. It can find a grammar of best fit from the ambiguous input by learning which combinations of parameter values are more probable than others.

The use of universal defaults more generally has been adopted recently as a fundamental component of the deterministic learning model in Sakas and Fodor (2012). This is perhaps the most developed work on identifying unambiguous evidence, attempting to do so for a large number of parameters in a vast domain of over 3,000 constructed languages. This is an enormous task, and the findings are compelling given the rich variety of languages and the sophisticated parameter space. With so much diversity, ambiguity in the input abounds and often results in many languages not having any unambiguous evidence for particular parameters. Having recourse to a set of universal default values thus becomes indispensible in ensuring that the learner does not converge on a non-target grammar.

Equipping a deterministic learner with a universal set of default values thus appears to give it some traction with the issue of the poverty of unambiguous evidence. Nevertheless, defaults offer no account for variability across learners of the same language. How could children (and thus adults) acquiring the same language end up with slightly different grammars, differing perhaps in the setting of only a single parameter P_1? If multiple learners are presented with unambiguous evidence concerning the value of P_1, then they will set it to the same value. If unambiguous evidence is sufficiently and systematically lacking, then they will all stick to the default value. There is no way for the learning outcome to diverge across children. Yet we will see again in Chapter 4 that such variability exists in Korean for a parameter for which it appears there is no (or vanishingly rare) unambiguous evidence. As will be discussed further, it is precisely because there is such a high degree of ambiguity in the input that such variability is

possible. Thus a desideratum for a learning model is that it be sensitive to ambiguous input in order to model such variability. The learning model I propose presents a proof-of-concept illustration of just such variability.

Finally, a universal set of defaults is insufficient to account for the full range of attested child errors. Suppose that at any point in time the child's grammar can be characterized by (a) the set of all parameters that have been set correctly on the basis of unambiguous evidence; and (b) the set of default values for all parameters that have not yet been set. Now let us consider a single parameter P_1 with values x and y. If a child is acquiring a language with a grammar that has the setting $P_1(x)$, then any errors of the form $P_1(y)$ must result from a default value of y for P_1. This is because there is not any unambiguous evidence in the language for $P_1(y)$; all the relevant unambiguous evidence would be evidence for $P_1(x)$. In this way, a deterministic model can, in principle, account for any given child's error: the error results from default parameter value.

However, if learners make use of a universal set of default values, then we only expect to see one kind of error in children cross-linguistically, i.e. an error that conforms to the default value. More precisely, the following kind of error is not predicted by a deterministic learner. Suppose Language_1 (L_1) has parameter setting $P_1(x)$, and that Language_2 (L_2) has parameter setting $P_1(y)$. Further suppose that the default value for P_1 is y. In this scenario, we predict that errors are possible only in L_1. This error would involve the child using the default value y for P_1. In contrast, the child acquiring L_2 will not use x for P_1: the child encounters no evidence for $P_1(x)$, and since the default value of y for P_1 is consistent with a grammar set to $P_1(y)$, the child's utterances will always appear consistent with the target grammar in the relevant respect. Nevertheless, there is evidence that this kind of error in children is attested. In Chapter 6, I present a more detailed discussion of English and Swedish that illustrates these contrasting error patterns. Whereas English has $P_1(x)$, Swedish has $P_1(y)$. However, there is evidence that children acquiring English produce utterances with value y for P_1, whereas children learning Swedish produce utterances with value x for P_1. No matter what the default for P_1 is, an error pattern is attested that cannot be accounted for under the deterministic learner we have been considering.

Such contrasting error patterns can be accounted for under a probabilistic model that learns from ambiguous input. In the model I propose, any parameter value (even non-target values) that is consistent with the input can be reinforced. If a sufficiently large proportion of the input is ambiguous with respect to a given parameter, the learner can be misled into temporarily adopting a non-target parameter value. This is especially true when the ambiguous input closely resembles some other non-target grammar, and the non-target parameter value appears to be a highly probable choice. In the case of the contrasting error patterns, we can say that some English input sufficiently resembles Swedish, and that some Swedish input sufficiently resembles English, such that a child acquiring one language can be temporarily misled into adopting a non-target grammar

that looks somewhat like the other. Under this view, ambiguous evidence plays a crucial role in accounting for a child's errors.

I conclude that a deterministic learner, even when augmented with a universal set of default values, falls short of addressing the acquisition phenomena of learner errors and variability across learners. An approach that relies exclusively on unambiguous evidence appears to be inadequate, and I have suggested that we can account for these learning phenomena with a learning model that crucially learns from both ambiguous and unambiguous evidence.

Having introduced in general terms what is at stake regarding ambiguous evidence, in the remainder of Section 2 I present a fairly schematic overview of what this ambiguous input could look like. I will also sketch out how the model I propose can learn from this ambiguous input. I describe two kinds of scenarios involving ambiguous evidence that I focus on in this thesis. The first is the general case of ambiguity in the input. The second is a more specific case of ambiguity and can be called the subset case. There is no deeply significant difference between these two scenarios. They both present a similar puzzle: how can this ambiguous evidence be informative for the learner? Moreover, the same learning model can provide a unified account of this puzzle and the various case studies I look at. A core contribution of this work is to address this puzzle by showing how the model learns in a systematic way from this evidence. Separating ambiguous evidence into two scenarios simply provides a clear way of highlighting how different aspects of the same model can learn from ambiguous evidence in ways that are independent of each other. Partly to this end, the presentation of the general scenario will involve ambiguous evidence that cannot be categorized as belonging to the subset scenario. A key component in my answer to the puzzle of the general scenario involves a novel insight about how parameter interaction can affect learning outcomes. As we will see in Section 2.2, even when the input is ambiguous, parameters can interact in such a way that certain combinations of values are favored as a function of the shape of the input. These values are thus more likely to be reinforced, and this results in the learner being systematically pushed toward adopting certain grammars. In contrast, parameter interaction need not at all be relevant for the subset scenario, in which case the model will rely on learning from implicit negative evidence.

I lay out schematic examples of the two scenarios below. After presenting each scenario, I briefly discuss the basic insight of how a probabilistic model could learn from the ambiguous evidence. Short previews of empirical examples are also given. These are taken up in more detail in the case studies in Chapters 3-5.

2.2 The general case

In this section I describe the general case of how the learner's input can be ambiguous with respect to multiple grammars. Suppose the learner is trying to decide between two hypotheses about what the adult grammar looks. One is that the phrase YP in the adult grammar is head-initial (1a), and the other is that YP is head-final (1b). Now suppose that the learner hears the string [XY], in which Y heads the complement of X. This string is an example of an ambiguous datum: it is compatible with either grammar in which Y's complement follows or precedes Y.

(1) *Input: XY*
a. Y-initial: $[_{XP}$ X $[_{YP}$ Y $[_{ZP}$ Z]]]
b. Y-final: $[_{XP}$ X $[_{YP}$ $[_{ZP}$ Z] Y]]

What is the learner to do in this scenario? Both grammars in (1) appear to be equivalent in capturing the simpler datum of [XY]. A novel insight of this work is that input that is similarly ambiguous can be highly informative to a probabilistic learner. To see how the model would learn from this input, though, it is necessary to consider a different example in which we crucially see how parameters interact with each other in a multi-parameter hypothesis space.

To see how the probabilistic model learns from parameter interaction, consider the following constructed example that illustrates the basic concept. In (2) there is a simple 2-parameter hypothesis space, which is presented schematically. There is a set of sentences S, which forms the input to the learner. There are also two binary parameters A and B, whose values can be set either positively or negatively. There are thus 4 logically possible grammars in the hypothesis space. If S is ambiguous, then there will be multiple grammars that are compatible with it. But if only three grammars are compatible with S, then some parameter values will be more likely to be compatible with S than other parameter values. For example, in (2) there is a parameter interaction such that A can only be positively set if B is positively set, and B can only be negatively set if A is negatively set:

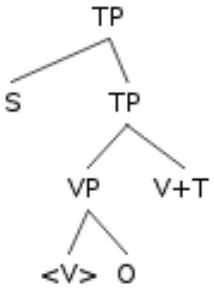
(2) *Hypothetical Learning Scenario I: 2-parameter grammar space*
a. Input: {S}
b. Grammars compatible with input:
1. [+A, +B]
2. [–A, +B]
3. [–A, –B]

Such parameter interaction means that of the input-compatible grammars, a majority of them (two-thirds) are [–A], and a majority of them (also two-thirds) are [+B]. By hypothesis, the grammar of best fit given S is [–A, +B]. The probabilistic learner is

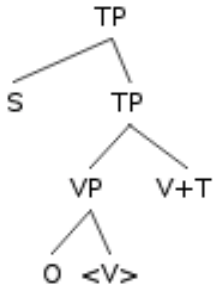
sensitive to this proportional tendency: on average the learner will expect a [–A] grammar, and on average the learner will expect a [+B] grammar. The grammar most likely to be compatible with the input, then, is [–A, +B], and this is the grammar that on average the learner's expectations will push the learner toward. Recall that the model generates sentences that correspond to the input. In terms of generating sentences, the learner is most likely to generate sentences within the set S if the learner chooses [–A] and if the learner chooses [+B]. On average, then, the learner will reinforce the weights/expectations of these values more and will thereby be pushed toward learning a [–A, +B] grammar. Thus parameter interaction can inform the learner as to what the most compatible grammar is, and in this way the model can learn in a systematic and principled way from ambiguous input.

In Chapters 3 and 4 I consider natural language examples of this sort in more complex parameter spaces. The case study in Chapter 4 looks at parameter interaction in Korean, where the entire corpus of input is ambiguous, as it is in (2). Korean is canonically verb-final, but there is ambiguity as to whether all syntactic projections are head-final. For example, the verb could vacate the VP by moving to T, in which case T is head-final, but then V could be either head-initial (3a) or head-final (3a'). Or V could remain in-situ, necessitating a head-final VP, but allowing for T to be either head-initial (3b) or head-final (3b'). This is possible if T undergoes some post-syntactic process of attaching to V such as Marantz's (1988) Morphological Merger, or any rule of affix lowering (cf. Chomsky 1957; 1981). This is illustrated schematically in (3) for a simple transitive clause with subject and object; the base position of a constituent is indicated with angled brackets.

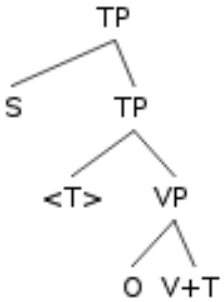
(3) a. *V-initial Korean*



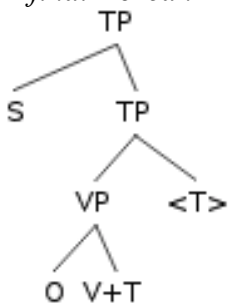
a'. *V-final Korean*



b. *T-initial Korean*



b'. *T-final Korean*



Despite this high degree of ambiguity, the most compatible grammar is one in which all its heads are phrase-final. By learning probabilistically from the effects of parameter interaction, I give a proof-of-concept illustration of how the model learns a consistently head-final grammar.

The other case study in Chapter 3 looks at parameter interaction in Swiss German. There we will see that even though the entire corpus of input is compatible with only one grammar, there is sufficient ambiguity in a large subset of the input for ambiguous evidence to play a significant role in the learning process. In particular, I propose that parameter interaction can cause children to initially learn a non-target parameter setting. I preview this case study in Section 3 below, where I take up the issue of modeling learner errors.

2.3 The subset case

The subset case presents a more specific example of the puzzle of learning the target grammar from ambiguous evidence. It also highlights how a probabilistic learner can resolve this puzzle by learning from implicit negative evidence when all the overt evidence is ambiguous. Before proceeding, I note that learning from implicit negative evidence and parameter interaction need not be mutually exclusive. Rather I present these two kinds of learning separately to show how each can independently serve as an effective method of learning, as well as an insightful way of modeling learners' performance. Both methods of learning fall out from the design of the model, which is to learn the grammar of best fit given the input.

A learning scenario involving subset and superset languages has figured prominently in discussions of learnability and language acquisition since at least the work of Gold (1967) and Wexler and Manzini (1987), although works such as Atkinson (2001) and Frank and Kapur (1996) have raised the possibility that syntactic parameters instantiating the subset case might not actually exist in natural language. In what follows I show that it is relatively straightforward to illustrate an example of the subset case when we consider different hypotheses for what the complement of a syntactic head might be. After presenting this schematically, an empirical example will be seen from zero-derived causatives in English, which builds on Pylkkänen (2008). Although I focus on how a probabilistic learner can address this particular example, the logic of how the learning mechanism works is of a general nature and has the potential to be applied in other possible empirical examples illustrating the subset scenario.

This work builds on the research program involving learning from implicit negative evidence in Regier and Gahl (2004), Hsu and Griffiths (2009), and Perfors et al. (2010). As regards a syntactic perspective, this work is related to the line of research concerning anaphoric *one* in English, initiated in Regier and Gahl (2004) and developed in a series of papers by Lisa Pearl (Pearl and Lidz 2009; Pearl and Mis 2011; Pearl and Mis *in press*).

These works on anaphoric *one* look at the intersection of implicit negative evidence, learning from ambiguous input, and syntactic representations. As discussed in Chapter 5, though, there are a number of differences between them and the current study. Perhaps most significantly for the discussion here is the recognition that indirect positive evidence plays a crucial role in learning anaphoric *one* (e.g. Pearl and Mis 2011). Indirect positive evidence is a particular kind of inference. By way of example, in Pearl and Mis the learner takes positive evidence about the properties of pronominals such as *it*, and uses that evidence to infer properties of *one*. The evidence for *one* is indirect, but it is based on positive evidence that the learner receives as input. In contrast, we will see that implicit negative evidence is sufficient for addressing the learning challenge with zero-derived causatives. Thus the present study serves to shine a light on the role of implicit negative evidence for the learner.

The subset scenario can be characterized as follows. Suppose the learner of a language must choose between two grammars G_1 and G_2 on the basis of the set of utterances of that language U_L. The set U_L can be defined as the set of string-meaning pairs that are possible in the language. Next, let us consider the sets of string-meaning pairs that can be generated by the two grammars in relation to U_L and to each other. In this scenario, both grammars can generate the entire set U_L. In the case of G_1, it generates exactly the string-meaning pairs in U_L and no more. However, in the case of G_2, the set U_L is a proper subset of the set of string-meaning pairs that can be generated by G_2. Thus the set of string-meaning pairs generated by G_1 is a proper subset of that of G_2, and we can call G_1 the grammar of the subset language, and G_2 the grammar of the superset language. This means that G_2 generates all the utterances that are possible in the language, but it also generates utterances that are not possible in the language. We can also infer that G_1 is the target grammar for the learner, as it is the grammar of best fit: it is the more compatible grammar in that the learner will not overgeneralize and make errors by adopting that grammar. But now we are faced with a learning puzzle. How can we be sure that the learner will settle on G_1 when both grammars are compatible with all the utterances of the language? The positive evidence available to the learner underdetermines the correct structural analysis, and both grammars are seemingly good options. The advantage of a probabilistic learner is that it is sensitive to the possibility of G_2 generating utterances that are not attested. This is implicit negative evidence, and the fact that these utterances are not attested means that from the learner's perspective, G_2 is less likely to be the target grammar.

A schema of the kind of example I consider for the case study in Chapter 5 is given in (4). Suppose there is some syntactic head X, and that there is a parameter governing what the complement of X can be. The complement could simply be YP, as in (4a), or it could be ZP, which properly contains YP (4b). Further, suppose that the target setting in the adult grammar is the simpler structure in (4a), but that all the evidence the learner receives is ambiguous – it is compatible with either the structure in (4a) or (4b). The

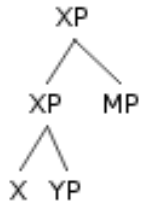
puzzle is then how the learner can be sure to learn the structure in (4a), when the structure in (1b) is seemingly just as good of a candidate.

(4) a. $[_{XP}$ X $[_{YP}$ Y …]]
b. $[_{XP}$ X $[_{ZP}$ Z $[_{YP}$ Y …]]]

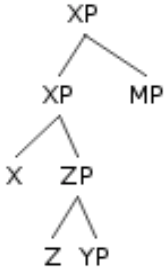
One possible way in which such ambiguity could arise is as follows. Let us consider the sets of string-meaning pairs that the two grammars could generate. If Z is phonologically null, and if there are no truth conditional or pragmatic differences between (4a) and (4b), then the string-meaning pairs of (4a) and (4b) are equivalent.

We have not yet seen how one language may be considered a subset of the other. The implications of choosing YP or ZP as X's complement can be augmented by considering the modification possibilities of adverbials. If there is an asymmetry in the modification possibilities of (4a) and (4b), then it becomes possible for the two grammars to diverge. Suppose there is some modifier MP that can adjoin to XP and ZP, but not to YP. This is schematized in (5):

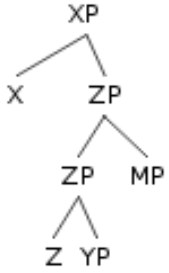
(5) a. *Subset language* (cf. (4a))



b. *Superset language* (cf. (4b))



b'. *Superset language* (cf. (4b))



Although the strings generated by the structures in (5) will also be identical, that is not necessarily the case with their interpretations. If MP adjoins to XP, as in (5a) or (5b), then both grammars will produce a string with modification of XP and thus the same interpretation. However, in the more complex structure it is possible to adjoin MP to ZP, as in (5b'), in which case XP is not modified, and a different interpretation is possible. In sum, the structures in (5a) and (5b) can be mapped to one string-meaning pair and are compatible with either grammar, whereas the structure in (5b') can be mapped to another string-meaning pair and is compatible only with the more complex grammar.

All else being equal, the set of string-meaning pairs generated by the simpler YP-complement grammar is a proper subset of those generated by the more complex ZP-complement grammar.[2] Now consider the possibility of a language that has no grammatical utterances corresponding to the string-meaning pair in (5b'). The string-meaning pair represented by (5b') was the one point of difference that was singled out for the two grammars. In such a language, then, all the grammatical utterances would be compatible with either parameter setting for X's complement. Nevertheless, from an analytical perspective, we can point to the absence of the utterances corresponding to (5b') as evidence that the speakers of this language have learned the subset grammar. Further evidence comes from speakers' judgments concerning the impossibility of these utterances. Such utterances are impossible in the grammar of the subset language; given that such utterances are not attested, we can conclude that the grammar is the subset one.

An empirical example of (4) involving zero-derived causatives from English is discussed in Chapter 5 and is illustrated below. The learning challenge stems from examples such as (6). Of two potential interpretations concerning modification of the adverbial in the causative sentence (6a), only one is possible. The unavailable reading is one in which the internal argument of 'awake' is characterized by grumpiness; this reading is available in the non-causative (6b).

(6) a. John awoke Bill grumpily.
✓John is grumpy (high reading)
✕Bill is grumpy (low reading)

b. Bill awoke grumpily.

According to Pylkkänen's (2008) analysis of (6), the unavailable reading is only possible in a structure like (4b)/(5b'). Pylkkänen concludes that given this unavailability, the structure of (6a) is like the structure of the subset grammar in (4a)/(5a).

But things are less clear for the learner. When a child is confronted with an utterance containing modification of XP (which can be determined by the interpretation), he or she could map such an utterance seemingly equally well to the grammar of either the subset or superset language. Indeed, there is no utterance in the learner's input that can serve as unambiguous evidence in favor of either grammar. In the face of such ambiguity, it is not clear how a population of learners would all learn the target grammar, namely the grammar of the subset language. In this sense, the subset scenario appears to present us with a case of the Poverty of the Stimulus (cf. Chomsky 1980). Given that the learner receives no overt evidence pertaining to the correct parameter setting, the question is then whether this parameter setting is unlearnable. Nevertheless, under the

[2] I note that what is crucial here for the subset/superset relationship is the sets of string-meaning pairs that the different grammars can generate. The fact that the structures of the superset grammar in (5b, b') do not embed the structure in (5a) does not bear on this relationship.

approach pursued here, the grammar of the subset language is learnable and thus no poverty of the stimulus arises.

My approach to this puzzle is to capitalize on the absence in the learner's input of the string-meaning pair schematized in (5b'), i.e. the absence in the input of the low reading in (6a). Such an absence is implicit negative evidence and can be highly informative for a probabilistic learner. This approach to learning incorporates insights from Bayesian learning. As Regier and Gahl (2004) discuss, an intuitive non-linguistic example of how this works is found in Laplace's (1825) law of succession. This is used to calculate the likelihood that the sun will rise tomorrow. Suppose there are two hypotheses: Hypothesis A is that the sun will rise every day, and Hypothesis B is that there is only a 50% chance that it will. The observed evidence is that the sun has risen every day for a great number of days. It is important to note that both hypotheses are consistent with this evidence. Under Hypothesis B we still expect to see the sun rise. But we also expect to see the sun not rise as well. This expectation is never realized. The fact that the sun has not yet failed to rise can be taken as implicit negative evidence, which can then be used to determine that Hypothesis B is less likely than Hypothesis A to be correct.

Let us return to the linguistic example from above. Each hypothesis about the data that the learner makes is accompanied by an auxiliary set of expectations about what the input will look like. Thus under both grammars there is an expectation that some proportion of the time the input to the learner will contain an utterance in which MP modifies XP. As the two grammars are largely identical, the expectations under each grammar are largely the same as well. Crucially, although all the auxiliary expectations given the grammar of the subset language can also be found under the grammar of the superset language, if the learner chooses the grammar of the superset language, there is an additional expectation. This additional expectation is that some proportion of the input will contain an utterance in which MP modifies ZP (and not XP). In contrast, under the grammar of the subset language there is no such expectation. Given that input to the learner does not contain such an utterance, the expectations of choosing the grammar of the subset language match the observed data better than those of the grammar of the superset language. Recall that the model is generating sentences that correspond to attested input. Given that there is an extra expectation under the grammar of the superset language, under that grammar the learner will sometimes generate sentences in which MP modifies ZP, as in (5b). But such a sentence is not attested in the input corpus. In contrast, no such unattested sentences are generated under the grammar of the subset language. All else being equal, this means that choosing the grammar of the superset language is not as good a match to the input as choosing the grammar of the subset language. Accordingly, on average the learner will reinforce the grammar of the subset language more because it is more likely to generate target sentences. Over time, as the observed data continues to repeatedly better match the grammar of the subset language,

the learner will gradually assign more and more probabilistic weight to choosing YP as X's complement. In this way the learning model is able to learn from the ambiguous evidence and arrive at the parameter setting for the simpler structure found in the grammar of the subset language.

In Chapter 5 I provide a proof of concept illustration of how to model learning the grammar of the subset language for causatives as described above. Under the grammar of the superset language, the learning model expects to hear input containing the unattested reading in (6a), whereas there is no such expectation under the grammar of the subset language. Given that the input does not contain this low reading, the model will be pushed toward learning the grammar of the subset language.

In the subset example above, a key component of the model's learning from implicit negative evidence is an asymmetry in the learner's expectations under the two grammars. In other words, the expectations given the grammar of the subset language are a proper subset of the expectations under the grammar of superset language. It is the additional expectation(s) of the superset grammar that provide the source for the implicit negative evidence and that push the learner toward the subset grammar. Importantly, it is the additional layer of structure, ZP in (4b), that we can point to as giving rise to this imbalance in expectations.

Before closing this section we can observe that the example from the previous section does not instantiate the subset case. Recall from (1), repeated below, that the parameter under consideration involved the precedence relation between a head and its complement.

(1) a. Y-initial: $[_{XP}$ X $[_{YP}$ Y $[_{ZP}$ Z]]]
b. Y-final: $[_{XP}$ X $[_{YP}$ $[_{ZP}$ Z] Y]]

As discussed in Safir (1987) and Atkinson (2001), a Y-initial grammar and a Y-final grammar can each generate a string-meaning pair that the other cannot, the different utterances corresponding to the different structures in (1). Thus neither grammar can generates a language that can be considered a proper subset or superset of the other. This holds for all the simple cases of parameter setting considered in Chapters 3 and 4, which look at parameter interaction. All the parameters in those chapters involve head-complement order or verb movement; none of the languages that result from grammars comprised of these parameters is a proper subset of any of the others. In contrast to the subset example in (4), in the more general example from (1) we are considering competing grammars that have the same inventory of syntactic heads with the same hierarchical relations. Although there are differences across the grammars in (1), this does not lead to an imbalance in expectations. Recall that such an imbalance was crucial in the discussion above concerning implicit negative evidence. Thus given that the expectations are symmetric under the different grammars in (1), there is no role for implicit negative evidence to play in learning the parameter setting for YP-headedness.

As all the grammars in the cases involving parameter interaction discussed below are not in any subset/superset relations, they all have such symmetric expectations. Accordingly, implicit negative evidence will not figure in any of the case studies that look at parameter interaction.

3. Modeling learner errors and variability

In Section 2, I discussed two empirical issues relating to language acquisition that learning from ambiguous evidence can help us understand: learner errors and variability in grammars. A core contribution of this work is to show how a probabilistic learner can be effective in shedding light on these phenomena and in modeling learners' performance. In this section I give a brief overview of the case studies related to these phenomena in Swiss German and Korean, which I take up in more detail in Chapters 3 and 4 respectively.

Much research has shown that throughout the developmental course of first language acquisition, the syntax of children's utterances is highly adult-like in its parameter settings (Wexler 1998; Snyder 2007). Nevertheless, these same studies have observed that there are corners of the grammar where children produce errors. Thus we can ask the following question is: why do children make the errors that they do? The answer to this question is likely to be multi-faceted. My approach is to build on the insight of Yang (2002) to show that one aspect of child errors is learning from ambiguous evidence.[3] In particular, I show that these errors can result from the effects of parameter interaction among competing grammars when learning from ambiguous input. As in (2) above, the learner will be pushed toward the most compatible grammar. However, this time the grammar of best fit is a non-target grammar.

In Chapter 3, I discuss Schönenberger's (2001, 2008) study of Swiss German children learning the position of the finite verb in embedded clauses. Schönenberger shows that the children pass through a developmental stage when they appear to consistently place the finite verb in the wrong position in embedded clauses. (7a) is an adult utterance, in which the finite verb appears in a clause final position. (7b) is a non-target child production, in which the finite verb follows the complementizer, preceding the other constituents in the embedded clause. Throughout, I indicate non-target child productions with #.

[3] Thus it is entirely possible that some errors are the result of maturational considerations (cf. Rizzi 1993/1994; Wexler 1998). For example, some parameters might be inactive early in development, only gradually emerging as an active part of the hypothesis space later in the learner's development. There is nothing in the learning model I propose that is incompatible with such maturational considerations, and in Chapter 2 (note 4) I briefly speculate as to how they could be incorporated in the model. As the null hypothesis, though, I assume that no such maturational considerations are operational in the case studies I investigate. Indeed, we shall see that this null hypothesis has success in modeling the acquisition phenomena at hand, including child errors.

(7) a. [Wenn t'Hex före-chunt]…
when the-witch outside-comes
'When the witch comes outside…'

b. # [Wenn chunt t'Hex före]… (M: 3;10)
when comes the-witch outside

(Schönenberger: 82)

The learning model here provides a proof-of-concept example showing that in the learner's initial state, parameter interaction can push the learner to a non-target grammar that is capable of generating the errors. This can happen when a large proportion of the input is sufficiently ambiguous. In Swiss German, there is a high degree of ambiguity as to the exact structural position of the finite verb. This ambiguity is especially prevalent in matrix clauses, which form the bulk of the input to the learner. The effect of parameter interaction on learning from this input is that a grammar with non-target verb placement is most likely to be compatible with the input. As the model is an input-driven learner, it will temporarily favor a non-target grammar that leads to errors. In effect, parameter interaction leads the learner to 'mis-set' a parameter about where the verb will be moved to. The influence of matrix clauses, which comprise a preponderance of the linguistic evidence, thus provides a natural and principled way of accounting for errors.

Moreover, the learning model is able to recover from this non-target state and learn the adult grammar. As the child progressively learns more and more target parameter settings, the target grammar emerges as the grammar of best fit. This is a welcome result that models the children's speech productions and accounts for their errors in a principled way through the learning process itself.

The second issue concerns grammatical variability. This is when we see evidence of learners making use of multiple grammars. How can we model such variability, and what are the circumstances that allow such variability to occur? Again, the answers here are more complex than can be addressed here; my goal is simply to show (a) the role that ambiguous evidence plays, and (b) how the probabilistic model can model this variability.

There are two aspects of variability that I will focus on. The first involves *within* speaker variation. We see this, for example, in another developmental stage of the Swiss children described by Schönenberger. In this stage, the children alternate between sometimes placing the finite verb in a target position in embedded clauses and sometimes placing it in a non-target position. At any particular point throughout this stage, the children have variable target and non-target productions. And measured over blocs of time, we see that there are varying proportions of target and non-target productions, with the children gradually approaching adult-like performance by the end of the study. This kind of variability is thus based on the learner making errors. Given the discussion above about the source of these errors, we see the importance of ambiguous evidence in giving rise to this variability. When we look at modeling this phenomenon, as Yang (2002)

discusses, the notion of grammar competition embedded in a probabilistic learner provides us with a natural link to capturing this variability. In the model proposed here, when presented with new input the learner's expectations change only gradually as probabilistic weights are shifted from one hypothesis to another. A consequence of this is that unless there is a single grammar that is considered to be overwhelmingly likely to match the adult grammar, there will be multiple grammars at the learner's disposal for practical use. We can model which grammar a speaker might use to produce an utterance on the basis of these weights. The heavier a weight is for a particular parameter value, the more likely the learner is to use that value when speaking. Consequently, weights that are more evenly distributed across a parameter's values result in more variability across productions. At any point in time, a learner's variable errors result from there being sufficiently heavy weights on competing parameter values. Further, we can model the variability over time between errors and target productions with the gradual shifting of weights from parameter settings of the non-target grammar to those of the target grammar.

The second type of variability I look at involves parametric variation *across* speakers in a population. This type of variability describes a scenario in which a population of language learners is presented with (approximately) the same corpus of input, but who nevertheless learn slightly different grammars. Each individual speaker's grammar is (relatively) stable, but across the population their grammars split into two or more groups. To see how this might arise, we again need to consider a corpus of ambiguous input as in (2), but this time there crucially needs to a be a lack of parameter interaction. This is shown schematically in (8).

(8) *Hypothetical Learning Scenario II: 2-parameter grammar space*
  a. Input: {S}
  b. Grammars compatible with input:
    1. [+A, +B]
    2. [–A, +B]

In (8), there are only two grammars compatible with the input. A compatible grammar must be [+B], but neither a [+A] or a [–A] grammar is favored: they both represent exactly half of the input-compatible grammars. As the probabilistic learner is not heavily swayed via parameter interaction to learn one particular grammar, the learner is 'free' as it were to learn either of the two grammars in (8b). So long as the learner is not too tentative in reinforcing one of the grammars, given sufficient time one of the grammars will emerge as the end state for the learner. In Chapter 4, I present a more complex proof-of-concept example of this concerning verb raising in Korean. Again there is an entire corpus that is compatible with multiple grammars, and because there is insufficient parameter interaction, the model ends up learning two grammars – one with and one

without verb raising. These results find support from the recent experimental work on Korean in Han et al. (2007).

In a sense, (8) brings us full circle to the subset example in (4). In (4) we considered one parameter setting that distinguished two grammars, with all the input being compatible with either grammar. (8) simply looks like a more explicit multi-parameter illustration of this. However, there are two important points to make about this type of variability. The first concerns implicit negative evidence. In the subset case it was possible for the model to learn from implicit negative evidence so as to systematically learn the same grammar. I have assumed that this is not always possible in non-proper-subset cases such as (8), and indeed in cases of variability across speakers, we have evidence that it is not. The second is an issue of complexity. The variability we see in (8) can be replicated in a more complex hypothesis space in which the competing input-compatible grammars differ by more than just a single parameter value. I consider an example of this in Chapter 4, where what we see is effectively the synthesis of (2) and (8): some parameter values are systematically learned by the model, whereas other values are learned variably across different runs of the model.

In sum, a probabilistic learner is well-suited to modeling both learner errors and grammatical variability. Further, in the examples pursued in the case studies here, it is ambiguous evidence that plays a crucial role in giving rise to both errors and variability.

4. Ambiguity and development

In this section I provide a brief overview of how ambiguity relates to various learning outcomes for the child, as well as the developmental trajectory for the child. In particular there is a connection between the amount of ambiguity and where we expect to see variability within a single learner and across learners. This section summarizes some of the points from earlier in this chapter and is meant to lay out concisely the different possibilities for variability that are illustrated by the different case studies that form the core of this work.

First, we expect to see some variability during the process. A core aspect of the learning model is that when presented with an ambiguous data point, there is some indeterminancy as to which grammar the learner will reinforce. As we have seen above, there is often a grammar(s) that is more likely to be reinforced, but that grammar will not always be reinforced given such ambiguous input. Learning can thus be characterized by a back and forth process as the learner shifts probability mass across competing grammars. How much back and forth there is depends on the amount of ambiguity in the input. If the input corpus contains a healthy amount of ambiguous evidence, then we expect to see variability during the course of acquisition as the child shifts expectations back and forth. This is what we see in Swiss German in Chapter 3. The pervasive ambiguity in the input allows the child to move back and forth between different hypotheses, alternating between target and non-target utterances. This variable

production continues until the learner converges on a single grammar that accounts for all the input.

Second, we expect to see some variability in the end-states of different learners. That is, given equivalent input, some learners adopt one grammar, whereas other learners adopt a minimally different grammar. Let us consider how this could happen. If the input corpus contained unambiguous evidence such that only one grammar was compatible with all the evidence, then we would not expect learners to arrive at different grammars. Indeed, the Swiss German children end up learning the same grammar because there is only one grammar compatible with the all the input. But what if multiple grammars were compatible with all the evidence? In such a scenario it is possible for learners to diverge and adopt different grammars. So long as learners are not too tentative with reinforcing parameter values on the basis of ambiguous evidence, it is then possible for one learner to adopt a grammar with one value for a given parameter, while another learner adopts a different value for that same parameter. Thus we see variability across speakers, but minimal variability within speakers. This is what we see in Korean in Chapter 4. In Korean, virtually the entire corpus is ambiguous with respect to verb movement, and we see that some learners adopt a grammar with verb raising, while others do not.

Full corpus ambiguity (i.e. a high degree of ambiguity across the entire corpus) is thus a necessary condition for variability in the end-states of different learners, but it is not a sufficient condition. There are multiple ways in which learners can arrive at the same grammar, even with full-scale ambiguity. One is parameter interaction. In Korean, there is insufficient parameter interaction such that neither a verb raising or non-verb raising grammar is overly favored. Both grammars are viable options. However, a high degree of parameter interaction can push all learners toward the same end-state. Further, implicit negative evidence can result in one grammar being favored over another grammar. In the case study on English zero-derived causatives in Chapter 5, we see that the grammar of the subset language is more compatible than the grammar of the superset language. The model always learns the simpler grammar in this case, and thus there is only a single learning outcome.[4]

[4] There is another noteworthy case in which we see syntactic parametric variability, but one that I will not explore in this work. This case involves substantial variability within speakers as a component of the adult grammar. Such variability is a signature characteristic of languages undergoing diachronic change. An example of this kind of variability comes from Old English regarding OV and VO word order (cf. Pintzuk 2002), which has been modeled by Pearl (2007). The variability in Old English can be characterized by what we could call mixed input. That is, there are opposing kinds of unambiguous input, each of which can be taken as unambiguous evidence in favor of grammars with parameter values that are opposite those unambiguously supported by other input. We can make sense of this variability from a modeling perspective: as the different strains of unambiguous evidence pull the learner in multiple directions, there are multiple grammars whose parameter values receive sufficient weight so as to remain viable options for the speaker. However, this is not the kind of input that we see in the case studies considered in this work, which looks primarily at the effects of ambiguous evidence. The study of within speaker variability during the course of diachronic change goes beyond the scope of this work, and it remains an open question as to whether all such cases can be modeled by means of what I have called mixed input.

To conclude, we can use the model and its relationship with ambiguous evidence to more fully describe variability in language acquisition. During the developmental course of acquisition we see variability as the learner assesses the fit of different competing grammars to evidence in the surrounding linguistic environment. This is not always the case, though, in the adult grammar. We also see variability with regard to learners' end-states. This occurs when we have a sufficiently great enough degree of ambiguity in an entire corpus such that a lack of parameter interaction leaves the learning outcome variable. This discussion has not been intended to provide an exhaustive classification of all possible learning trajectories and outcomes, but what it does do is situate ambiguous evidence squarely in the center of these kinds of considerations.

Chapter 2

The Learning Model

1. Introduction

In this chapter I present a more detailed description of the probabilistic learning model that I propose to address the learning scenarios that were introduced in the previous chapter. Although a full treatment of how the model works in these different scenarios is reserved for the empirical discussion of them in the case studies of Chapters 3-5, the introduction of the model in Section 2 shows generally how the model addresses these scenarios. I illustrate this by means of two toy examples in Section 2.

I present two versions of the model in Section 2. The full version is a *generative* model. By 'generative' I have in mind the general conception of generative syntax, whereby the model builds syntactic structures and accounts for all the utterances of any particular language. As we shall see, the model does so by using different grammars to generate sentences. Suppose the model is trying to learn a particular language *L* and is exposed to utterances from *L*. The grammar that has the greatest likelihood of generating the sentences that are actually found in a given language is the grammar that the model will end up learning. The objective of a generative learner is to maximize the likelihood of a particular data set (the utterances it has been exposed to) given a particular choice in grammar.

A generative model contrasts with what can be called *discriminative* models (which I discuss in more detail in Section 3). A discriminative model may not or may learn from ambiguous evidence, as in Sakas and Fodor (2001) or Yang (2002) respectively. What crucially characterizes such a model is that it merely checks the compatibility of a given input datum with a particular grammar. If the input is incompatible with that grammar, then the model will move the learner away from that non-target grammar; if the input is compatible with that grammar, the model will keep or try to keep the learner at that grammar. The objective of such a learner is to end the learning process in the target state. In a discriminative model it is sufficient to simply end up in the right state, and the model is silent on how a sentence is generated.

The subset learning scenario is thus a challenge for a discriminative learner because for such a learner the grammars of both the subset and superset languages are compatible with all input available to the learner. A discriminative learner has no way of systematically favoring one of the grammars over the other. In contrast, under a generative model the learner ends up in the right state by considering how sentences are constructed. A generative model assesses the overall fit of a grammar to the input and settles on the grammar that best matches the input. As regards the challenge of the subset scenario, it is the grammar of the subset language that has a greater likelihood of

generating the sentences that comprise the corpus of input found in the subset language. A generative model thus provides a deeper theory of the relationship between inputs and grammars, and this allows such a model to be a better approach to modeling language learning.

To expediently provide proof-of-concept demonstrations for some of the case studies, it will often be sufficient to use a simplified version of the model, which is also discriminative. I describe this simplified implementation in Section 2, as well as how this simplification can be more expedient. The fully generative implementation will be crucial in addressing the subset learning scenario, and I will use the simplified implementation for the non-proper-subset learning scenarios in Chapters 3 and 4. Still, it is to be understood that the fully generative implementation is intended to be used even in non-proper-subset scenarios and that we expect comparable results for these scenarios under both versions of the model.

In Section 3, I compare the model to three alternative models for learning that have been proposed in the literature. These models are all discriminative learning models. What differentiates these models is that they broadly represent different approaches to the role that ambiguous evidence plays in the learning process. On one end of the spectrum, a model such as that in Sakas and Fodor (2001) tries to do all learning without ambiguous evidence. On the other end is the probabilistic learner in Yang (2002), which learns from both ambiguous and unambiguous evidence. Yang's model shares strong parallels with the simplified version of the model I propose, although I have already pointed out that this simplified implementation is embedded in a more complex generative model, which differs from Yang. As discussed in Section 3.3, a further difference with Yang's model concerns the nature of the probabilistic weights used by the learner. Overall we shall see that these alternative models have varying degrees of success in addressing the learning challenges from Chapter 1, but not without limitations. A strength of the model I propose is that is provides a unified account of the different learning scenarios under consideration.

2. Overview of the model

In this section, I present a general introduction to the kind of Bayesian-inspired model I will be using for language learning and parameter setting. In Section 2.1, I demonstrate the basics of how the model works by using two toy examples to walk through some core aspects of the learning procedure. Doing so allows us to see upfront the basic way in which the model can respond to the challenges of ambiguous evidence. A concise but also more complete summary of the learning procedure is reserved for Section 2.2. The interested reader may wish at some point to jump ahead to this summary before returning to Section 2.1. In the interest of exposition, though, I have put a more general introduction to the model first in Section 2.1, which walks more slowly through various steps of the learning procedure. Section 2.1 also introduces both implementations of the

model, the fully generative one and the simplified version. I discuss when it is sufficient to use the simplified version of the model, namely when the question of learning from implicit negative evidence is not at issue. The simplified version of the model will allow for a more streamlined and expedient approach to modeling Korean and Swiss German in later chapters.

The focus of this research is to look at how the model learns from ambiguous evidence, but this is not to say that unambiguous evidence plays no role in the learning process. Indeed, unambiguous evidence plays a vital role in learning the target grammar. In Section 2.3 I situate the two kinds of evidence with respect to each other by comparing the different ways they affect how the model learns. Finally, more technical discussion concerning probability distributions and how probabilities are updated can be found in Section 2.4.

2.1 Introducing the model with two toy examples

In this section, I walk through the core aspects of the learning procedure with two toy examples. These examples are a useful way of introducing how the model responds to the challenges of ambiguous evidence. The first of these toy examples illustrates the subset learning scenario, in which the model learns from implicit negative evidence. The second illustrates the more general learning scenario of learning from ambiguous evidence, in which the competing grammars do not generate any languages that are in a proper subset relation with each other. In this more general scenario, the model is not learning from implicit negative evidence.

Consider the toy example in (1). There is a simple corpus of input to learn from (1a), and a highly restricted grammar to learn, i.e. a grammar with only a single parameter, which concerns head-complement order in the VP (1b). Suppose that UG makes available three ways to set the parameter for VP-headedness (1c): V could be obligatorily head-initial or obligatorily head-final, or there could be optionality, in which case both SVO and SOV would be possible.

(1) *Toy example #1*

a. Input corpus:	1. SV 2. SVO
b. Parameter:	1. VP-headedness: Does V precede or follow its complement?
c. Hypothesis space:	H1. Language is always V-init(ial) H2. Language is always V-fin(al) H3. Language is optionally V-init/V-fin

Because SOV is not possible, the target grammar is instantiated by H1. However, the possibility of optionality in (1c) now presents the learner with a subset problem. The

subset language contains SVO utterances, while the superset language contains SVO and SOV utterances. Given infinite input of the form SV(O), a discriminative learner can conclude that head-finality is not obligatory for V and easily eliminate H2. SVO is simply not compatible with H2. But how can the learner conclude that optionality is not possible? Both H1 and H3 are compatible with all the input. Proposing a V-initial default (which the learner could abandon if encountering SOV input) is not especially helpful here. After all, we can flip the learning problem by slightly changing the input corpus. Suppose there is another language with only S(O)V input – how could we preclude the learner from overgeneralizing and allowing SVO as possible? We could say that there is a V-final default, but then we must abandon the claim that there is a V-initial default and would be no better off with the original subset problem from (1). Moreover, proposing a V-initial or V-final default leads to the expectation that all learners pass through a VO or OV stage respectively, a claim that finds no empirical support (cf. Atkinson 2001).

Although (1) presents a learning challenge, there is a basic insight that will give us a handle on this difficulty. The learner will continue to hear SVO but never hear SOV. The insight is that the continued lack of SOV is too great a coincidence to be left to chance. The likelihood of the corpus in (1a) is greater given an obligatory V-initial grammar. The absence of SOV is implicit negative evidence according to which the grammar of H1 is the grammar of best fit.

How can we implement this insight in a learning model? The model I propose will assign a probability to each sentence; the probability will differ depending on which grammar is chosen. During the course of learning, the learner will be pushed toward the grammar with the highest probability for the sentences found in the input.

Let us see more concretely how this implementation will work by assigning a probability to the SVO input from (1). I now walk through many of the basic steps of the learning procedure (see also Section 2.2 for a fuller summary of the learning procedure). The learning process begins with the learner receiving a token of input, here an SVO utterance. The learner will then make a number of choices to generate a sentence (string-meaning pair) that matches this SVO input. This sentence of output is not an actual utterance produced by the learner. Rather, the process of generating output can be used as a metric for determining how well different hypotheses fit the evidence in the surrounding linguistic environment. The learner will generate a sentence by choosing a grammar. Choosing a grammar consists (minimally) of choosing a set of parameter values (which are familiar from the framework of Principles and Parameters; cf. Chomsky 1981, 1986), as well as different phrase structure rules. These structure-building choices can be associated with compositional meanings (cf. Heim and Kratzer 1998), with the full set of choices for each token of output giving rise to a particular meaning for that sentence. However, in this work I largely abstract away from semantic interpretation, instead assuming an idealized learner that has adult-like compositional semantics. All of these choices are associated with some probability. Thus the learner can

be described as a 'probabilistic generative model', and making choices can be called 'sampling'. Further, as the null hypothesis, I assume that there are no combinatoric constraints on what choices can be made for any given sample (although such constraints can stipulated; see Section 7 of Chapter for an example of how this might be implemented). Also as the null hypothesis, I assume that learning one parameter value does not intrinsically bias the learner toward learning some other parameter value (though I consider revising this position in Chapter 6).

Let us now look at the choices necessary to generate SVO output.[1] In particular, let us consider how to generate a VP with VO order, as it is the VP that is most relevant for setting the parameter in (1b). Relevant choices are schematized in (2). (2a) shows the different choices concerning the parameter values from (1c). I assume that at the beginning of learning, the prior probabilities associated with these choices are all equal, say .333.[2] If the learner chooses optionality (2a.3), then there is a subsequent choice to determine what the outcome of that optionality will be for this particular sentence. We can note, then, that the model can be organized such that some choices are conditional on having first made some other choice.[3] Having made a choice concerning what the structure of the VP is, the learner can next make a choice as to whether an object will appear in the sentence output. Only certain choices (in boldface below) lead to generating a sentence that matches the input; we can call such a sentence 'target output'.[4]

---

[1] Thus the model, insofar as it is characterized by all the choices discussed here, can be represented by a probabilistic context-free grammar (cf. Perfors et al. 2006).

[2] For the purpose of exposition, in this section I use simple values such as .5 as priors. This will be revised in Section 2.4 where I introduce a somewhat different notion of what the weights are that the learner assigns to different parameter values, and how these weights are used to generate what the learner thinks is the probability of a given parameter value in the adult grammar. Nothing in the discussion in this section hinges on this distinction, though.

[3] The model thus has the possibility of having some parameters be hierarchically dependent on others, although as the null hypothesis I will assume that no parameter is hierarchically dependent on any other parameter. This lack of hierarchy is implicit in a number acquisition models such as Gibson and Wexler (1994) and Yang (2002) (though see Baker 2001 for a detailed hierarchy of major syntactic parameters based on typological claims).

[4] To determine whether some output is target output, all the model would need is some (possibly domain general) comparison metric that allows it to evaluate whether a sentence of output is the same as a sentence of input. I abstract away from how this is done, instead assuming an idealization of the learner that is able to evaluate this matching perfectly. (Part of this idealization, then, is that the learner understands the meaning of each input sentence and can correctly identify each schematic element in the input, such as S, O, Adv, etc.) In Chapter 1 (note 3) I noted that the model can in principle be modified so as to include a maturational view of development (cf. Rizzi 1993/1994; Wexler 1998). For example, some parameters might be inactive early in development, only gradually emerging as an active part of the hypothesis space later in the learner's development. Thus under a maturational view, early in development the generative choices available to the learner could be such that for some tokens of input the learner would not be able to generate output that matched perfectly. It could still be possible to learn from such input. What would be needed is a more refined evaluation metric. According to this new metric, output that was sufficiently close (according to some criteria) to matching the input could be considered target output, and parameters could be reinforced accordingly. As none of the simulations I run implement any such maturational considerations on learning, I leave the development of a more sophisticated evaluation metric for future research, and will make use of the idealized evaluation metric mentioned above.

(2) *Generating VO output*

a. VP-headedness?

1. V-init:

S VP → S V (O) ***p* = .333**

2. V-fin:

S VP → S (O) V *p* = .333

3. Optionality ***p* = .333**

i. V-init: **S VP → S V (O)** ***p* = .5**
ii. V-fin: S VP → S (O) V *p* = .5

b. Is there an object?

1. If Oblig. V-init or Optionally V-init:

i. **S V (O) → S V O** ***p* = .5**
ii. S V (O) → S V *p* = .5

2. If Oblig. V-fin or Optionally V-fin:

i. S (O) V → S O V *p* = .5
ii. S (O) V → S V *p* = .5

The likelihood of generating target VO output in (2) can be summarized as follows. The probability of VO output given obligatory V-initial is .5. In contrast, the probability of VO output given obligatory V-final is 0. Obligatory V-fin is clearly not a grammar that matches the input well. What about optionality? If the learner chooses optionality, there are two subsequent choices that must be made to generate VO output: the choice for V-initial (*p* = .5) in (2a.3i), and that there is an object (*p* = .5). Given optionality, the joint probability of these two subsequent choices is .25. The conclusion is this: the likelihood of generating target output is greater under obligatory V-initial.

A procedural note is in order here. What happens if the learner generates a sentence that does not match the input (i.e. non-target output), such as SV or SOV? In that case, the learner will sample, or make choices, again until target output is generated.

Suppose the learner has generated target output. In that case, the learner will equally reinforce the weights associated with the choices made to generate that target output. Using the schema in (2), we can implement this by bumping up the probabilities associated with these choices (see Section 2.4 for a revised implementation of the reinforcement of weights). We saw that generating target output is most likely to happen under an obligatory V-initial grammar. As an example, if the learner generated target output by choosing obligatory V-initial, then the reinforcement could look as in (3), with the probability associated with obligatory V-initial increased by some amount α. When a choice is reinforced, the likelihood of choosing it again is increased because its associated probabilistic weight has increased. The purpose of reinforcement is to maximize the likelihood of making similar choices (which led to target output) given similar input.

(3) *Reinforcing choices*

a. VP-headedness?

1. V-init:

S VP → S V (O) $p = .333 + \alpha$

2. V-fin:

S VP → S (O) V $p = .333 - \alpha$

3. Optionality $p = .333 - \alpha$

i. V-init: S VP → S V (O) $p = .5$
ii. V-fin: S VP → S (O) V $p = .5$

b. Is there an object?

1. If Oblig. V-init or Optionally V-init:

i. **S V (O) → S V O** $p = .5 + \alpha$
ii. S V (O) → S V $p = .5 - \alpha$

2. If Oblig. V-fin or Optionally V-fin:

i. S (O) V → S O V $p = .5$
ii. S (O) V → S V $p = .5$

The procedure I have just outlined will iterate with all subsequent input until the end of the learning process. On average the learner will reinforce obligatory V-initial more than the other parameter values. Given sufficient SVO input and sufficient reinforcement of obligatory V-initial, the learner will be pushed toward a parameter setting of obligatory V-initial. Parameter setting happens when the weight for a given value is very strong; in the example here, the probability for that value would approach 1. This learning procedure allows the model to learn from implicit negative evidence and thus captures the insight that an obligatory V-initial grammar is the grammar of best fit.

One way of thinking about the probabilities involved in making choices is that they represent expectations on the part of the learner about what the input from the primary linguistic data will look like. These expectations then affect what the model outputs. In the example in (2), the learner would initially expect roughly a 50-50 split between OV and VO orders in the input and would initially generate output that reflects that chance distribution. In actuality, the split is 100-0 in favor of a VO order. As a learner, the child wants to maximize the probability that the output generated matches the input. The learner can do this by updating the prior probability with a posterior probability that more accurately reflects the distribution of the input. Posterior probabilities are updated only incrementally. Accordingly, the learner's expectations change gradually as well. The input thus has a conditioning effect on the learner. This conditioning effect can be thought of as the learner's gradually modifying his/her expectations so that they are more informed by the surrounding linguistic environment. In the example from (1), over time the probability mass will have shifted so much in favor of obligatory V-initial, that the learner's expectation for a V-final utterance will be vanishingly rare. At this point we can say that the learner has set the parameter as V-initial.

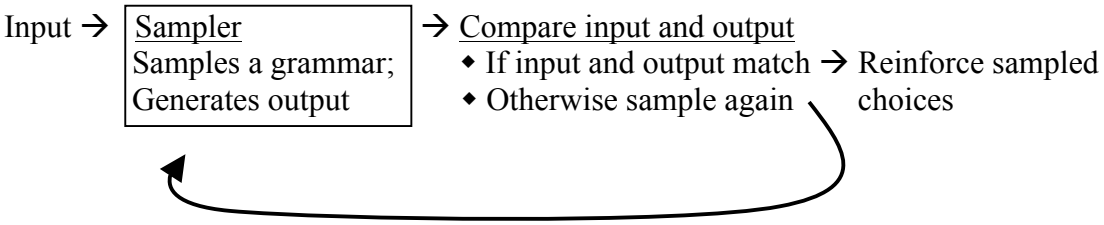
The probabilities associated with a parameter's values also have a straightforward link to the utterances a child actually says. The model can use the probability distributions to determine different properties of a learner's spoken productions. Given

that these probabilities shift gradually, we can thus account for learner productions that are both non-target and variable. For example, suppose that the updated probabilities are such that the likelihood of making a choice with V-initial is .6, and the likelihood of making a choice for V-final is .4. As the learner has not yet finished learning that the target grammar is V-initial, we thus expect learner errors of VPs with an OV order. As the learner's expectation is that V-fin reflects the adult grammar only 40% of the time, we expect the learner's non-target productions to vary with target ones according to the parameter values' probabilistic weights.

In sum, the learning process can be characterized by a competition of different hypotheses concerning different parameter values. Different samplings of parameter values represent different grammars. The nature of this competition is the online updating of the probabilistic weights associated with the different parameter values. How well a grammar fares in this competition depends on how a good a fit it is with the input. The probabilistic weights associated with a grammar's parameter values, then, are a direct measure of how well the grammar fits the input.

I have just outlined the basics of the learning procedure for the generative model, which is the full version of the model. This is illustrated schematically in Figure 2.1 below.

Figure 2.1 The Generative Learning Model



Let us now consider a second toy example. This will help to become familiarized with the learning procedure and will help to set the stage for the introduction of the simplified implementation of the learning model. This second toy example is also representative of the more general learning scenario with ambiguous evidence, in which the competing grammars under consideration do not generate languages that are in a proper subset relation with each other.

Consider the second toy example in (4). In (4a) there is a fairly simple input corpus, but now there is a contrast between VO and OV orders in matrix and embedded clauses respectively. The focus of this toy example will be to learn the structure of the VP, a task that is challenging because of the contrasting word order patterns in different clauses. This contrast is reminiscent of word order patterns that we see in German and previews to a certain extent what we will see in the case study of Swiss German in Chapter 3. The grammar to be learned is also small; this time there are two parameters, one for VP-headedness, as well as a generic parameter concerning verb raising. For simplicity, I will

abstract away from the possibility of optionality regarding VP-headedness. I will also abstract away from the position that the verb could raise to (an issue that I discuss in great detail in Chapter 3). For our purposes here, it is sufficient that if the verb raises out of the VP under a [+Raising] grammar, then head-complement order within the VP is underdetermined. Assuming that the object remains within the VP, and again abstracting away from where the verb raises to, the important observation here is that [+Raising] is compatible with both matrix and embedded clauses.

(4) *Toy example #2*

a. Input corpus:	1. Matrix: SVO 2. Embedded: …[SOV]
b. Parameters:	1. VP-headedness: Does V precede or follow its complement? 2. Verb raising [±Raising]: Does V move out of the VP?

The example in (4) immediately presents a learning challenge owing to the ambiguity of the input. The challenge is that each individual token of input is compatible with multiple grammars. Let us look more closely at this ambiguity by considering how a learner might respond to the input. Suppose the learner hears matrix SVO input and thinks that the target grammar is [–Raising, V-init], which is compatible with SVO. Then what happens when the learner hears embedded OV. Embedded OV is not compatible with the learner's hypothesis of V-init. The learner could adjust one of the parameter values. But which value does the learner flip? A raising grammar could help: a [+Raising, V-init] grammar is compatible with embedded OV. But V-final could also work: a [–Raising, V-fin] is also compatible with embedded OV. Suppose the learner adopts a [–Raising, V-fin] grammar. This will work in the short-term, but then what happens when the learner next hears matrix SVO again? The learner could toggle back and forth between the values V-initial and V-final. We certainly don't want to say that there is optionality with respect to VO/OV because matrix SOV is unattested. What I have just described is a fairly haphazard learning process, which underscores the challenge of learning from ambiguous evidence.

Let us take a step back to consider a way to account for what happens in both types of clauses in (4) with the same grammar (i.e. the same set of parameter values). A generative model gives us a principled way to learn the most compatible grammar for the entire corpus of input. The insight is that the likelihood of the entire corpus is greatest given a certain combination(s) of parameter values, namely a [+Raising, V-init/fin] grammar, which is the best fit to the input.

Walking through the learning procedure I described above allows us to see how the model can learn such a grammar of best fit. Recall that the model is trying to generate sentences that match the input, which is matrix VO and embedded OV. Table 2.1 presents a consolidated summary of the relevant choices and the output sentences these

choices generate. The leftmost column gives some sentence-specific choices concerning what kind of constituents will be present in the output (e.g. is there an object?). The other columns represent choices for different parameter values. A ✓ indicates the choices can generate target output, whereas a ✗ indicates the choices generate non-target output. What we see is that every grammar can generate sentences that match part of the input corpus, but only a [+Raising] grammar can generate target output for the entire input corpus.

Table 2.1 Schematic choices and sentences generated in the generative model for toy example #2

Choose a…	[+Raising, V-init/V-fin]	[–Raising, V-init]	[–Raising, V-fin]
matrix object	✓SVO	✓SVO	✗SOV
subord. CP with object	✓…[SOV]	✗…[SVO]	✓…[SOV]

Thus if the learner chooses a [+Raising] grammar, there is a greater likelihood of generating a target sentence. This means that on average the learner will incrementally reinforce [+Raising] more than other parameter values, thereby increasing the likelihood that the learner will choose [+Raising] given future input. Despite the pervasive ambiguity in the input, the learner will gradually be pushed toward a parameter setting of [+Raising], thereby learning a grammar of best fit.[5]

We have now seen how the generative model addresses the challenges of ambiguous evidence in the two toy examples. At this point, a simplification to the model becomes expedient. The reason for this is simply the great number of choices that need to be made when we scale up the complexity of the examples so as to try to model natural languages. In (5) I give the schematic corpus that I use for the simulation of Swiss German in Chapter 3. In (5) we see that there are a great number of choices that the learner would need to make in order to generate target output. For example, the learner must choose whether the output has an auxiliary, an embedded clause, an adverb, etc., and if there is an adverb, where it adjoins to, and so on.

(5) *Schematic corpus for Swiss German*

a. SV	e. SV[Comp. SV]	i. AdvVS	m. OVS
b. SVO	f. SV[Comp. SOV]	j. AdvVSO	n. OAuxSV
c. SAuxV	g. SV[Comp. SVAux]	k. AdvAuxSV	
d. SAuxOV	h. SV[Comp. SOVAux]	l. AdvAuxSOV	

The complexity of a corpus such as (5) means that there are many paths the learner could

[5] And as for VP-headedness, given the randomness of the sampling, a learner that is not too tentative could sufficiently reinforce either value to adopt either a V-initial or V-final grammar. Modeling a learner that is too tentative is discussed at various points in the remainder of this chapter.

take that lead to generating non-target output. For example, suppose the learner hears OVS output and generates an OVSAdv string. Did the learner get a core property of the grammar wrong, or did the learner get the grammar right but simply made the wrong sentence-specific choice and ended up generating the wrong sentence? As linguists, we might be inclined to conjecture that the learner did indeed make the right parametric choices, or at least a number of correct parametric choices. The model makes no distinction between these different possibilities. The trouble is that there is a vast search space among the network of choices to find the right set of choices to generate target output for any particular token of input. Recall that the model will keep sampling and generating output until a match with the input is reached. With so many choices, the sampling procedure could continue for quite some time, and this is not practical for running the proof-of-concept simulations I have in mind. Thus we have the question of how to proceed when dealing with a rich corpus of input.

The presence of a vast search space is not a problem *per se*. More sophisticated sampling procedures have certainly been developed to address the issue of rare event sampling in a vast search space. See Kratzer et al. (2014) for an example of a particular methodology of searching within a vast space. Nevertheless, the focus of this work is not on efficiently navigating through a large search space, and a fruitful approach for present purposes will be to abstract away from a number of sentence-specific choices as much as possible, while still maintaining the generality of the learning approach.

As a simplification to the model I propose the following. Generally it is sufficient to just sample values for the core parameters, such as head-complement order or whether there is verb raising, etc. The learner will then check to see whether these are compatible with a given token of input. Implementationally, this compatibility is encoded in the model.[6] Similar to the full version of the model, if these choices are input-compatible, then the learner equally reinforces these values, otherwise the learner samples again until a input-compatible set of values has been sampled. Again, this simplified learning process will iterate with subsequent tokens of input until the end of learning, with the learner gradually being pushed more and more strongly toward particular parameter values. In terms of the fully generative model, the learner is checking to see whether a set of parameter values is part of a larger set of choices that could lead to generating target out. That is, there is a non-zero probability of generating target output with these parameter values in the full version of the model.

The simplified version of the model that I have just described is in fact a discriminative learning model. In the introduction to this chapter I discussed how a generative model can be superior to a discriminative model. It is important to understand, then, that the use of a discriminative model here is intended only as an expedient way of

[6] This is a simple, albeit brute-force way of getting the simplified version of the model off the ground. The fully generative version of the model does not need to have any notion of compatibility hard-wired into it; all it needs is some general comparison metric, as mentioned in note 4.

illustrating certain kinds of learning results in some of the case studies. The expectation is that we would obtain comparable results under the fully generative version of the model. The question then becomes when can we use this simplified implementation? It certainly has the potential to be helpful when there is a large number of choices to be made in generating output. However, the number of choices cannot be used as the sole criterion in determining whether to use this simplification. The crucial point is whether or not the competing grammars under consideration generate any languages that are in a proper subset relation with each other. We saw such a relation in the first toy example in (1). The example in (1) illustrated the subset learning scenario, which I introduced in Chapter 1. In the subset scenario, it is insufficient to merely check for compatibility between the input and sets of parameter values because with a discriminative learner there will always be multiple sets of parameter values that are compatible with all the input. In such a scenario, a discriminative learner has no way of systematically favoring one grammar over another. In the subset scenario, it becomes necessary for the model to learn from implicit negative evidence (i.e. the absence of certain sentences in the input). This was implemented in the full version of the model by generating sentences. In (1), the grammar of the superset language was less likely to generate target output because sometimes it will be generating the unattested SOV sentences. The grammar of the subset language only generates sentences that are part of the subset language. It thus is more likely to generate target output and is a better match to the corpus of input. More generally, when there is no subset scenario, it is possible to use the simplified implementation of the model. Sets of parameter values that are input-compatible will be incrementally reinforced, and the learner will be systematically pushed away from sets of parameter values that are not input-compatible. Let us now consider some examples of using the simplified version of the model in cases that do not instantiate the subset scenario.

First, we can see that using the simplified implementation with the second toy example in (4) gives us comparable results. Now the learner samples only the core parameters of verb raising and VP-headedness. Table 2.2 provides a summary of these parametric choices and which kinds of input are compatible with those choices. A ✓ indicates the choices are compatible with some attested input, whereas a ✗ indicates the choices are compatible with some unattested input.

Table 2.2 Schematic choices and compatible input in the simplified model for toy example #2

[+Raising, V-init/V-fin]	[–Raising, V-init]	[–Raising, V-fin]
✓SVO	✓SVO	✗SOV
✓…[SOV]	✗…[SVO]	✓…[SOV]

Table 2.2 can be compared with Table 2.1. Table 2.2 is essentially an abbreviated version of Table 2.1: we are still finding the same grammars of best fit, but we are doing it with fewer choices. This is possible, crucially, because none of the grammars in Table 2.2

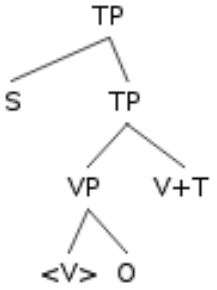
generates a language that is in a proper subset relation with any other language. This was already illustrated in Table 2.1 where we saw that for each set of parameter values there is some sentence that set of values cannot generate but which some other set(s) of values can.

I note that the absence of a proper subset relation in (4) holds (and thus differs from the first toy example in (1)) because a parametric choice for optionality with VP-headedness was excluded. Including a parameter value for optionality is not a general difficulty for the model, as we have already seen in how the generative model addresses the issue in the example in (1). In the interest of expedience, though, I will abstract away from a parameter value of optionality in the case studies developed here.[7] As optionality introduces the possibility of subset and superset languages, doing away with optionality will allow us to pursue the simplified version of the model in various case studies. More generally, if we consider competing grammars that do not have parameter values for optionality, but that differ in values for head-complement parameters (e.g. VP-headedness) and head-movement parameters (e.g. V-to-T movement), these grammars do not give rise to languages that are in a subset/superset relation with each other (cf. Safir 1987; Atkinson 2001).

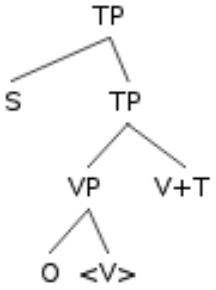
Before concluding this section, I sketch a different kind of example that makes use of the simplified version of the model in the general case (i.e. a non-proper subset scenario). Here I return to the case study of Korean that I first mentioned in Chapter 1, and which is discussed in more detail in Chapter 4. In contrast to the second toy example in (4), this example presents a different challenge in learning from ambiguous evidence: in Korean there are multiple grammars that are compatible with all the input. This example also allows us to focus on parameter interaction, which was introduced in Chapter 1. Parameter interaction concerns how certain combinations of parameter values are more likely as a function of the shape of the input. Recall from Chapter 1 that I claimed there are a number of grammars that are compatible with SOV input in Korean. In Chapter 4, I claim that these grammars are indeed compatible with (essentially) all the input in Korean. These grammars are schematized below in (6); all have different settings for VP-headedness, TP-headedness, and V-to-T movement.

[7] Further, preliminary research indicates that including a value for optionality does not present any obvious difficulty in obtaining results that are comparable to those reported in the case studies here.

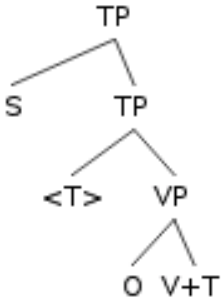
(6) a. *Verb raising Korean*



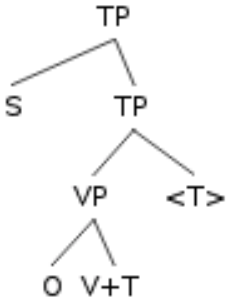
a'. *Verb raising Korean*



b. *Verb in-situ Korean*



b'. *Verb in-situ Korean*



Thus the structures in (6) illustrate how this input is ambiguous for every single parameter value of the 3 parameters mentioned above. This contrasts with the second toy example in (4). We saw in Table 2.2 that only one grammar was compatible with all the input. Thus (6) presents a different challenge in learning from ambiguous evidence: how can the learner systematically arrive at certain parameter values given that multiple grammars are compatible with all the input? First we can note that even though multiple grammars are compatible with all the input, these grammars do not result in subset/superset languages so long as we assume there is no parameter value for optionality, as discussed above. The case of Korean, then, is suitable for the simplified version of the model. And despite the high degree of ambiguity, parameter interaction has the effect of making ambiguous evidence informative for a probabilistic learner by pushing the learner to favor some parameter settings over others. What distinguishes the parameter space in (6) is that there is an interaction between verb raising and head-complement direction. Without verb raising, VP must be head-final (6b, b'), whereas with verb raising, TP must be head-final (6a, a'). In the simplified version of the model, the learner samples sets of parameter values that are compatible with the input (here SOV). The sets of input-compatible values for the parameters under discussion are represented in (6), and we can see (a) that a majority of the structures in (6) are V-final; and (b) that a majority of them are T-final.[8] All things being equal, this means that on

[8] In this example, the hypothesis space contains only the 3 parameters for VP-headedness, TP-headedness, and V-to-T movement. Ultimately, we would like to include in the hypothesis space all parameters that do interact with one another (leaving open the possibility of abstracting away from parameters that do not interact). Some parameters that are likely candidates for interacting have been excluded from this work,

average, the learner is more likely to choose V-final and more likely to choose T-final. With repeated input, these two values will be reinforced more. Thus the effect of parameter interaction is for the model to consistently learn a grammar that is both V-final and T-final, even though the input is ambiguous with respect to both values.

I note that this simplified version of the model shares at its core much of the learning mechanism in Yang's (2002) probabilistic model. Both models share the basic notion of checking to see whether a sampled grammar is compatible with the input. As will be discussed in more detail in Section 3.3, though, there are two points that distinguish the work here with Yang's. First, Yang does not explore the ramifications of parameter interaction, though they are predicted in his model. In contrast, I use the simplified version of the model to show how parameter interaction plays a crucial role in accounting for learner errors and variability in the case studies involving Korean and Swiss German. Moreover, as I have mentioned above, the simplified implementation is embedded in a more complex generative model that is intended to cover all the cases investigated here with the simplified version. Yang's model includes no component of generative capacity and thus is not able to address the puzzle of consistently learning the grammar of a subset language. The full version of the model I propose thus provides a unified approach for a range of different learning scenarios and empirical phenomena.

In summary, in this section I have introduced a generative learning model that is conditioned by what the input it receives looks like. The model samples parameter values and reinforces certain values in response to how well they fit the input. Reinforcement leads to greater probabilities for those values, which results in their being more likely to be sampled again and reinforced again. Sufficient reinforcement results in parameter setting. To expediently illustrate some of the proof-of-concept simulations, I introduced a simplification to the model, which simply looks at the compatibility of sampled parameter values with the input. As a general heuristic I will use the simplified implementation in the general case of learning from ambiguous evidence, and this is the approach I take in Chapters 3 and 4. Such an approach is insufficient for learning from ambiguous evidence in the subset scenario, and in Chapter 5 I run a simulation that uses the full generative version of the model.

---

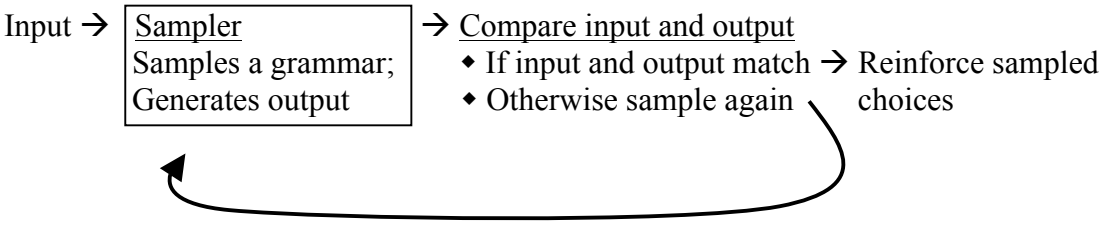
most notably parameters concerning the positions of the subject. The primary reason for doing so concerns the variability in the formulation of these parameters throughout the literature. I have instead focused on a core set of interacting parameters involving head-complement direction and verb movement that are formulated fairly consistently throughout the literature. This core set of parameters gives us a clear testing ground for the effects of parameter interaction on learning. The results are promising and provide a foundation on which to base future work. At various points in the discussion of the case studies in Chapters 3 and 4 I consider the possible effects of including additional interacting parameters. Although there are some complications, what emerges is a general picture that appears to be consistent with the findings in this work.

2.2 The learning procedure: A summary

Here I briefly summarize the basics of the learning procedure. This is largely a concise review of the procedure that was introduced in Section 2.1. However, an important component of the procedure that was not introduced in the previous section is introduced here. This is notion of *chews*, which concerns the number of times per token of input the sampling procedure is successful. As discussed below, varying the number of chews per token of input can be used to model learners that are more or less tentative, an important point to consider when learning from ambiguous evidence. The remainder of Section 2 will be spent fleshing out some of the details of the learning procedure summarized here.

I begin with the full version of the learning model, a schematic illustration of which is repeated in Figure 2.1 below. First, the learner receives a token of input. This input is an utterance (a string-meaning pair) from a corpus (which corresponds to the primary linguistic data). The learner will then attempt to generate output (a string-meaning pair) that matches the input. To generate output, the learner will construct a grammar by making a series of probabilistic choices. One thing the learner will sample is a set of parameter values. These parameters are of the sort familiar from the Principles and Parameters framework (Chomsky 1981, 1986). Each parameter value has a weight that is used to generate a probability; these weights represent the learner's expectations or beliefs about what the shape of the adult grammar is. The probabilities that are generated from these weights are then used to sample parameter values. In addition to sampling parameter values, the model will make an additional series of choices among different phrase structure rules concerning the shape of the output. For example, the learner can choose whether or not the verb will have an object, or whether the verb will be modified by an adverb. Thus the learner will be using the grammar it has sampled (i.e. the set of choices sampled) to generate a token of output. These structure-building choices can be associated with compositional meanings (cf. Heim and Kratzer 1998), with the full set of choices for each token of output giving rise to a particular meaning for that sentence. However, in this work I largely abstract away from semantic interpretation, instead assuming an idealized learner that has adult-like compositional semantics. Having generated some output, the learner next compares the output to the input. If they match, then the learner equally reinforces, or increases, the weights of the choices that were sampled to generate that output. The values of the weights after reinforcing (i.e. updating) them, are stored for future sampling. Detailed discussion regarding the nature of the weights and how they are reinforced can be found in Section 2.4. If the output does not match the input, then the learner repeats the sampling process until matching output is generated. With every subsequent token of input, this process repeats, the sampling being done with the most recently updated weights. By systematically reinforcing parameter values that lead to target output, the leaner can arrive at a grammar whose parameter values best reflect the shape of the input.

Figure 2.1 The Learning Model



The simplified version of the learning model is minimally different from the full version just described. This time when presented with a token of input, the learner will construct a grammar by sampling only a set of parameter values. Further, instead of generating some output, in this implementation, the learner will simply check to see whether the grammar sampled is consistent (or compatible) with the input. If it is compatible, then all the parameter values sampled will be equally reinforced. In terms of the full version, the grammar sampled has a non-zero probability of being able to generate matching output. As discussed in Section 2.1, the purpose of this simplification is to streamline the sampling process by forgoing a number of probabilistic choices that could lead to non-target output.

There is one other crucial procedural aspect of the sampling procedure that holds across all implementations of the model. We can model how tentative a learner is by varying the number of times the sampling procedure is successful per *token* of input. The model can be specified to generate target output a certain number of times per token of input. Thus the model can be required to go through the process of making choices so as to generate matching output 10 times, 100 times, etc. In the simplified version of the model this would entail sampling a compatible grammar however many times was specified. We can call each successful attempt to generate target output (or sample a compatible grammar) a 'chew'. It is possible that the model will make different choices for different chews of the same token of input. This happens when the input is ambiguous. Consider, for example, input that is an SV string. This input is ambiguous for the parameter concerning head-complement order in the VP, and the learner could generate target output by choosing V-initial on some chews and V-final on some other chews. In this case, both choices will be reinforced in proportion to how many chews involved each choice.

By varying the number of chews per token of input, it is possible to model how a learner may be more or less tentative. We can say that a learner is tentative if it assigns on average less weight to what might be a wrong hypothesis. When the evidence is unambiguous, the model will not make a choice that differs from the target grammar. However, when the evidence is ambiguous, any choice the model makes might be a non-target one. A tentative learner will try to avoid putting too much weight on any of these choices. In a sense, this learner is hedging its bets. One way of implementing this is by increasing the number of chews. All things being equal, then the greater the number of

chews, the greater the likelihood is that the learner's choices will be distributed equally across all the relevant hypotheses. This means that the different hypotheses will be reinforced equally. This is analogous to flipping a fair coin many times. The more times one flips the coin, the more likely half of the time it will land heads. If one flips the coin a small number of times, there is a greater chance that all the times the coin will land heads. Similarly, with only a small number of chews, there is a greater likelihood that only a single hypothesis will be chosen. That hypothesis would then be the only hypothesis that is reinforced for that token of input. If that hypothesis is a non-target one, it will then be harder for the non-tentative learner to recover from such an error. However, if a learner is too tentative, or conservative, it will be difficult to learn from ambiguous input: the learner will be too hesitant too assign much weight to any particular grammar. In the case studies I investigate, I claim that much of the learning is driven by ambiguous evidence. It thus is important to have a learning model that is not too conservative. Accordingly, the number of chews per datum in the various implementations of the model will be relatively low.

We have now seen an overview of the learning procedure. In the following sections I discuss in more detail some of the components of this procedure, including more precisely how the model reinforces parameter weights.

2.3 Ambiguous vs. unambiguous input

Unambiguous evidence has played a prominent role in developing a variety of learning models (cf. the discussion in Section 3), and indeed it is also important for the current learning model. This kind of evidence is input that for a given parameter is compatible only with a single value of that parameter. The focus of this work is to explore the role that ambiguous evidence plays in the learning process, but that does not mean all learning is to be attributed to ambiguous input. In the analysis of Swiss German in Chapter 3, unambiguous input ultimately plays a crucial role in helping the children move from a non-target grammar to the target grammar. As subsequent chapters will largely be concerned with ambiguous input, in this section I will briefly discuss learning from unambiguous input in the framework of the learning model. This will also help to put learning from ambiguous input into perspective by considering the contrast in how the model responds to both kinds of input. The case to focus on here is the non-proper subset learning scenario involving parameter interaction, given that there is no positive unambiguous evidence that can distinguish the grammar of subset language from that of the superset language. In broad terms I will contrast the effect unambiguous input has on the learner versus that of ambiguous input that pushes the learner toward a particular parameter setting via parameter interaction.

To explore the difference between the two kinds of input, recall from Section 2.3 that the learning procedure will have a pre-specified number of chews per token of input. In the full version of the model, this means that the learner will continue sampling until

output matching the input is generated, for example, 100 times. The choices that led to the matching output will then be reinforced in proportion to how many chews they were sampled for. Similarly, in the simplified version of the model, for each token of input the learner will continue to sample until grammars compatible with the input have been chosen the requisite number of times. Again, the parameter values of these grammars will be reinforced in proportion to how frequently they were chosen for these compatible grammars.

With this in mind, we can have an intuitive sense of how a little bit of unambiguous evidence can go a long way. First, every time the model is presented with a token of unambiguous input, we know certain parameter values that the model will reinforce. Further, these values will be reinforced for every chew of that input. This contrasts with ambiguous input: different parameter values can be reinforced on different chews, and this can result in only a gradual shift in a parameter's weights as the probability is dispersed across the different hypotheses. Unambiguous input thus has a more dramatic effect on shifting the probability mass associated with a parameter's weights. Because unambiguous evidence is so effective at shifting a parameter's weights, it can push the learner toward strongly favoring a particular parameter setting with a relatively small number of tokens of unambiguous input. As this learning is rapid, it is possible in principle to set a parameter with a relatively small proportion of the corpus comprising unambiguous evidence for one of that parameter's values. And the greater the proportion of the corpus that is unambiguous, the more rapid we predict this learning to be.

In the previous section I discussed how parameter interaction can help the model learn from ambiguous evidence. Some parameter values are more frequently attested in grammars that are compatible with the input. On average, these values will be sampled more frequently, and thus they will be reinforced more. The effectiveness of ambiguous evidence in pushing the learner toward a particular grammar can thus be attenuated in a way it cannot with unambiguous evidence. Suppose a setting of V-final is favored for a given type of input, given that a majority of the grammars compatible with that input are V-final. It is still possible that for any given token of that input, V-final might not be reinforced. Further, even if V-final is reinforced, it might not be reinforced for a majority of the chews; thus for that token of input the probabilistic weights will shift in favor of V-initial. What we do expect is that on average, for a given token of this input type V-final will be reinforced for a majority of the chews of this token. Learning from ambiguous input thus depends on this average. Not only must the type of input that favors V-final be a sufficiently large proportion of the corpus that the model learns from, but the margin by which V-final is favored given this input type must also be sufficiently large. The greater the proportion of compatible grammars that are V-final is, the more likely the learner is to be pushed toward adopting a V-final grammar. Cases where we expect to see parameter interaction playing a role in parameter setting thus involve the following two components: (i) a substantial proportion of the input is ambiguous and favors a particular

parameter value; and (ii) a sufficiently large majority of the grammars compatible with that input favor that parameter value.

Given the impact that learning from unambiguous evidence can have, the effect of learning from ambiguous evidence is often overshadowed. In the case studies of Korean and Swiss German we will see that the conditions are right for the role of ambiguous evidence to emerge to the forefront: there is an insufficient amount of unambiguous input, and certain parameter settings are heavily favored with ambiguous input in a sufficiently large proportion of the corpus.

2.4 Prior probabilities and the update procedure

In this section I present a more technical discussion of how parameter values are sampled during the learning process. The discussion will cover the form that the parameters' weights take, as well as how they are reinforced.

In Section 2.1, I simplistically assumed there were simple probabilities associated with the model choosing different parameter values. For example, with two values for VP-headedness (again excluding the possibility of optionality), the weights for each choice would initially be probabilities of .5. In the actual implementation of the model, instead of assuming that the weights are simple probabilities, I will assume that the weights are used to sample probabilities that are sampled from a probability distribution, a dirichlet distribution. One advantage of doing this is it removes an aspect of determinism from the model by allowing the model to randomly sample what the probability actually is. In other words, when the model is learning a particular parameter setting, we can suppose that the model doesn't knows what the actual probability is for any parameter setting. In a sense, that probability is what is trying to be learned. Instead we can model the learner's expectations as to what the probability of a parameter value is in the adult grammar. This is a way of capturing the learner's (un)certainty.

Taking into consideration the strength or weakness of a learner's certainty is a useful way capturing how a learner moves toward a conclusion when presented with different amounts of data. To take a simple example, suppose Mary is trying to figure out how likely her friend Susan is to make a free throw in basketball. If she sees Susan make 4 shots out of 5, she might have an expectation that the probability is most likely to be 80%. But how certain is she? After having observed only 5 shots, she might not be feeling especially confident, and might also expect, to a lesser degree, that the actual likelihood is closer to 70% or 90%. If overall she sees Susan make a total of 15 out of 20 shots, then with the additional observations, Mary can be more certain that the likelihood is closer to 75%.

Similarly, with equal priors there is the expectation at the outset that any one of the parameter values is just as plausible (as the correct analysis of the adult grammar) as any of the others. What the probability of each value is in the adult grammar is unknown. This probability can be sampled many times from a probability distribution and will vary

with each sampling. However, as there is no initial bias in the learner's expectations, the average of these probabilities (in the limit) will be 50-50 for a binary parameter. As the learner is exposed to more data, expectations will be revised by increasing the weights of certain parameter values to better reflect the nature of the input. And if a particular parameter value *x* has a much greater weight than the other value *y* for a binary parameter, the learner will be much more certain of a high probability for *x* in the adult grammar.

Learner expectations and certainty can be modeled with the dirichlet distribution. Using the dirichlet distribution to model varying degrees of (un)certainty has been applied in a number of domains and can be considered a psychologically plausible approach to learning (cf. Kemp et al. 2007). The dirichlet distribution can represent all the possible probabilities for a particular choice in the model, even when we don't know exactly what the probabilities are. The dirichlet distribution has *n* parameters, which are numbers > 0 called pseudo-counts, and each pseudo-count represents some expectation for a particular choice-point in the model. The pseudo-counts, then, are the weights corresponding to the expectations of different parameter values. In the example we started with, the choice-point concerned a V-initial or V-final grammar. This can be represented with a dirichlet distribution with two pseudo-counts: *dir*(*x*, *y*). With only two pseudo-counts, sampling from the dirichlet distribution will give us two probabilities (one for each choice of grammar) that sum to 1. Each time we sample, the probabilities can be different, but how different they are is a function of how certain the learner is about a particular value being attested in the adult grammar.

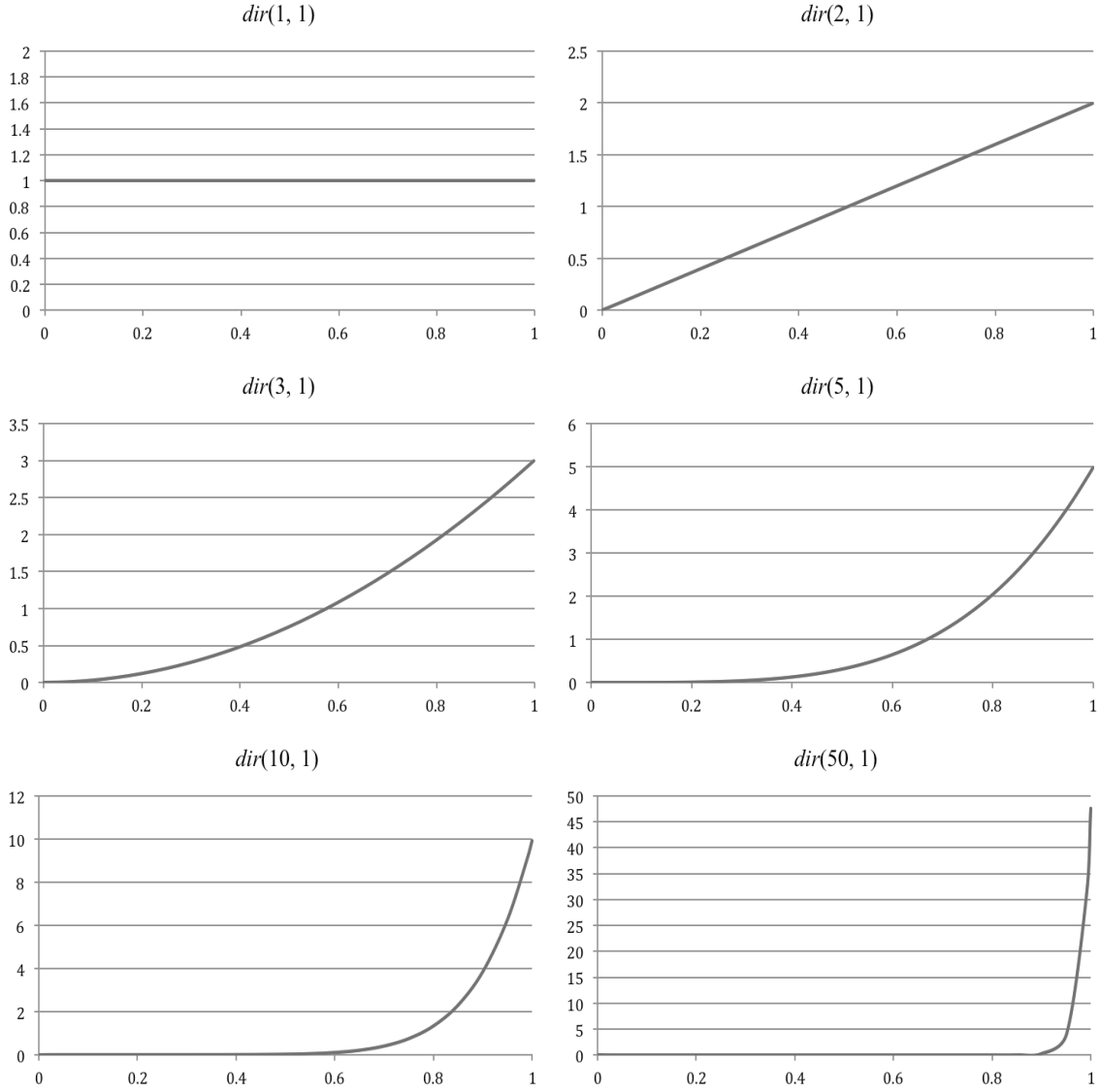
Another advantage of using a dirichlet distribution is that there is no upper bound on how many pseudo-counts there can be. Thus it is possible to expand the model to cover parameters that have more than two possible values. An example of this includes the full range of Pylkkänen's Cause-selection parameter (see Chapter 5) by including the possibility for a Phase-selecting setting; and this can be represented with a third pseudo-count: *dir*(*x*, *y*, *z*). A dirichlet distribution with three pseudo-counts will give us three probabilities that sum to 1.[9]

The confidence a learner has in its expectations can be illustrated by varying the values of the pseudo-counts (i.e. by varying parameter weights). An increase in value indicates an increase in certainty. I will discuss below more precisely how the learner updates expectations by increasing parameter weights in the face of linguistic evidence. For now, we can observe the changes in a learner's certainty graphically. Suppose we start with pseudo-counts of (1, 1) and progressively increase the weight for parameter value *x*. Examples of these new pseudo-count values are (2, 1); (3, 1); (5, 1); (10, 1); and (50, 1). Figure 2.2 shows the probability density function for these different pseudo-count

[9] The dirichlet distribution is the multi-dimensional generalization of the beta distribution. As all the parameters in the simulations here involve only two values, using the beta distribution would be equivalent to using the dirichlet distribution in this work.

totals. In the initial state, the learner is agnostic and is equally likely to sample any probability for value *x*. As we increase the pseudo-count value for *x*, though, the learner becomes more and more certain about what the probability for *x* is. By increasing the value for parameter value *x* to 2 or 3, we see a skew begin to emerge. An asymmetric preference is developing, such that the majority of the probability mass is now greater than .5. The learner is now moving toward a grammar that is more likely to have value *x* than value *y* for a particular parameter, but the learner is not especially certain of this: with weights (2, 1) or (3, 1) there is still a range of probabilities that the learner could sample. As more of a body of evidence emerges for value *x* (and not for value *y*), the skew in the probability mass becomes more and more pronounced. The majority of the probability mass becomes more tightly clumped around a range of probabilities that is closer to 1. With parameter weights of (50, 1) in favor of *x*, the learner is much more certain that the probability the adult grammar has value *x* is greater than .9.

Figure 2.2 Illustrating changes in a learner's expectations by varying parameters of the dirichlet distribution
(Graphs shows probability density function with given parameters; x-axis is probability; y-axis is probability density)



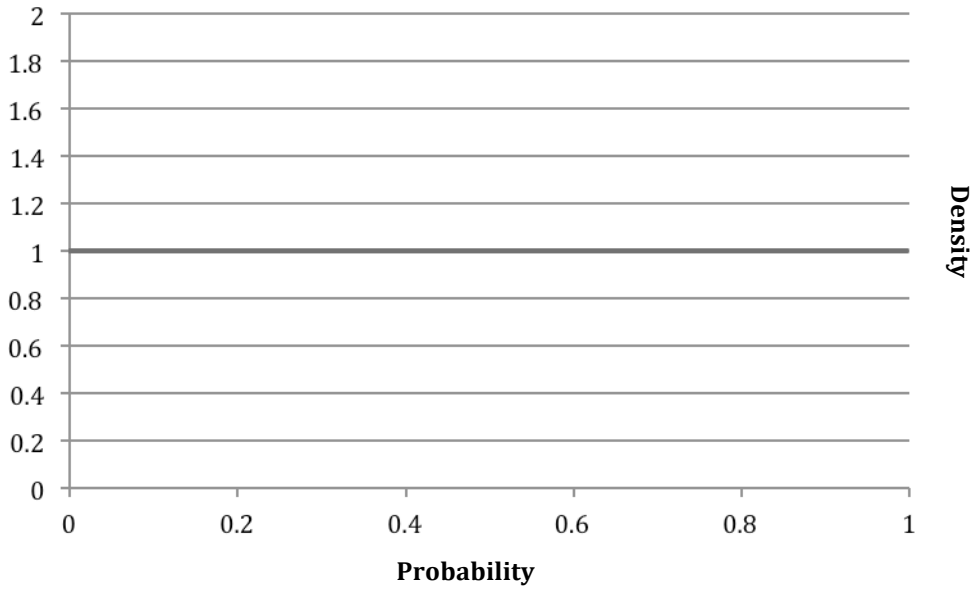
Having introduced the dirichlet distribution, I now discuss what the initial weights, or priors, are. As the null hypothesis, I will assume that the model begins learning with equal priors for all parameter values. For example, initially the choice between a V-initial or V-final grammar could have prior pseudo-counts of (1, 1). This means that there is an expectation that both grammars are equally likely, and that with infinite sampling from the dirichlet distribution each parameter value will be chosen 50% of the time (cf. the probability distribution with pseudo-counts (1, 1) in Figure 2.2). On any given sample, though, the probability of choosing say a V-initial grammar will be *p* within the range of [0, 1], and the probability of choosing a V-final grammar will be (1 –

p). In this way, the dirichlet distribution is able to represent the learner's expectations without specifying an exact prior probability.

As I have mentioned, the null hypothesis is that a parameter's values have equal priors, and I will adopt this as a general approach in the model's simulations. This is an idealization of the learner, in which the learner has no biases. It is unlikely that learners have no initial biases, but the empirical task is to first identify corners of the grammar where these biases may lie. To that end, the idealization of equal priors is helpful. With this idealization, we can push the insights of the learning model to their limit, and where they fall short, we have a candidate for a learning bias. The question I will be concerned with in this work is to see how much ground can be covered with the simple working assumption of equal priors. The case studies indeed show that a range of different learning scenarios and outcomes can be modeled without any initial learning bias. An additional and perhaps welcome result is that a candidate for a learning bias does emerge. In the discussion about how to expand the model for Korean in Chapter 4, I suggest that there might be a bias for a parameter concerning object movement. Possible support for this comes from additional production errors. Thus, adopting equal priors provides us with a useful starting point for modeling learning. They can be used to successfully model some phenomena, and they can help us probe areas where future investigation can reveal further evidence for biases.

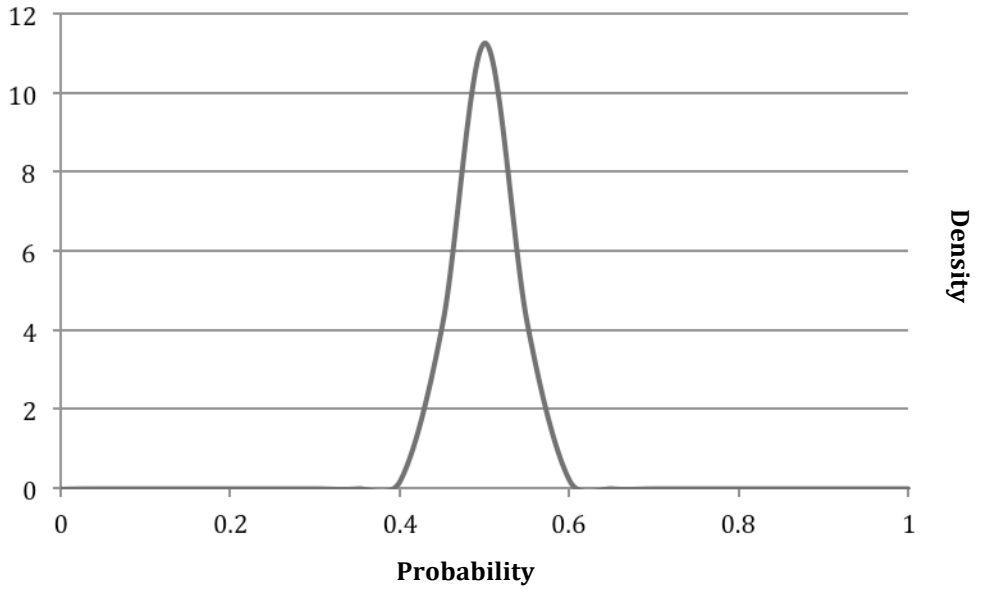
Having initial pseudo-counts of (1, 1) also allows the model to have weak, but equal priors. With equal pseudo-counts, the distribution is not skewed, and each grammar will be chosen approximately 50% of the time. Figure 2.3 below repeats from Figure 2.2 the probability distribution with pseudo-counts (1, 1). The distribution is uniform: on average the probabilities generated for each hypothesis will be equal, and each probability in the distribution is equally likely to be generated.

Figure 2.3 Dirichlet probability density function, parameters = [1, 1]



Importantly, with equal pseudo-counts that are small, there is only a weak expectation that no asymmetric preference will develop at an early stage of learning. An asymmetric preference in favor of a particular parameter value has already been illustrated in Figure 2.2. As we shall see in more detail below, an asymmetric preference emerges when one hypothesis is sampled more than the other, and this sampling preference can happen when the learner is more likely to have one hypothesis associated with a higher probability. To see in what sense a prior of (1, 1) is weak, consider the possibility of beginning the learning process with initial pseudo-counts of (100, 100), which would constitute a stronger expectation that there will be no asymmetric preference. The difference between starting to learn with pseudo-counts of (1, 1) and (100, 100) is that initially there is a higher likelihood with the former than the latter of sampling a probability that deviates more from 0.5. Thus with pseudo-counts of (1, 1) one is more likely to sample probabilities that are closer to 1 and 0 than with pseudo-counts of (100, 100). This can be seen by comparing Figure 2.3 with Figure 2.4. Figure 2.4 shows the probability distribution with pseudo-counts (100, 100). We see that although the distribution in Figure 2.4 is not skewed, the probability mass is tightly peaked around 0.5, meaning that the competing hypotheses are likely to both have a probability of 0.5 associated with them. As represented in Figure 2.4, the learner is strongly agnostic, and does not have much of an expectation that the probability of the parameter value *x* deviates much from .5 in the adult grammar.

Figure 2.4 Dirichlet probability density function, parameters = [100, 100]



In contrast, although the learner with priors of (1, 1) is agnostic about which parameter value is attested in the adult grammar, the learner is only weakly agnostic. This learner is more likely to explore the possibility that a parameter value has a high probability of being attested in the adult grammar. As there is a greater initial likelihood of generating higher probabilities with the learner in Figure 2.3, all things being equal, it is more likely that one hypothesis will be strongly favored early on in the learning process. This will be illustrated in more detail below when we consider the update procedure.

I will assume that the model begins learning with equal priors that represent weak expectations. This means that the weights for parameter values will initially be low, such as pseudo-counts of (1, 1). With low initial values, the learner is not strongly agnostic about the value of a parameter and can move relatively quickly toward strongly adopting one value over another. There are two related reasons why this is desirable. First, by having such a prior with weak expectations, the learner is less tentative: because the probabilities with pseudo-counts of (1, 1) can vary more, with finite sampling from the dirichlet distribution there is a greater likelihood for an asymmetric preference to develop, especially at an early stage of learning. We will see how such a preference can develop when discussing the updating procedure below. As was mentioned in the previous section, having a learner that is not too tentative is important when trying to learn from ambiguous evidence. If the input is ambiguous, but the learner is reluctant to put more weight on one hypothesis than the other, then it will be less likely for the learner to favor one hypothesis. The case of verb raising in Korean (see Chapter 4) is an important example showing how learners are not overly tentative: when faced with ambiguous input, different learners learn different parameter settings. A tentative learner is also less likely to make mistakes. The case of verb placement in child Swiss German

illustrates that learners do make mistakes. Thus we want a learning mechanism that is not too reluctant to put substantial weight on a non-target hypothesis on the basis of limited evidence, even if that hypothesis will ultimately be rejected. Priors with weak expectations, then, in addition to lower numbers of chews per input token (cf. Section 2.2), constitute a second way of modeling how tentative a learner is. A second reason for adopting priors with weak expectations is a more practical one, in line with this work being a proof-of-concept in nature. These priors allow for some hypotheses to be strongly favored more quickly than strong priors do. Again, we will see this more clearly when considering the update procedure below. Using priors with weak expectations thus provides an efficient way of showing how the current approach to learning works.

I now turn to how parameter weights are reinforced, that is, how expectations and certainty change. This is the updating procedure, which in its most basic form involves increasing the pseudo-count values. Recall that when a choice (such as a parameter value) that is sampled results in output matching the input (or in the simplified model, results in a grammar that is compatible with the input), the value of that choice is reinforced, changing the learner's expectations. The learner is trying to increase the likelihood of making a similar choice when encountering similar input in the future. As pseudo-counts represent the respective weights of the learner's expectations, we can indicate an increased expectation for a given choice (i.e. reinforcing a choice) by increasing the pseudo-count value for that choice. We can call pseudo-count totals that reflect changes in the learner's expectations updated values. Let us consider some examples of how to increase pseudo-counts before looking at how these updated values affect the likelihood of the learner making a particular choice. Suppose the learner is presented with a simple SV string, which is ambiguous for being V-initial or V-final. Further suppose the learner is not at all tentative and performs only one chew per input datum. If the learner samples V-initial and generates target output, then the pseudo-count total for V-initial will be increased. If the priors were totals of (1, 1), then the updated totals will be (2, 1) in favor of V-initial, with the weight for V-initial having been increased from 1 to 2. A more tentative learner will have multiple chews per token of input. This results in more iterations of sampling a successful grammar, and the results of these samples can be distributed equally across the values that are sampled. For SV input, if the learner has 100 chews with a sampling success rate of 47 times for V-initial and 53 times for V-final, then the updated pseudo-count totals will be (1.47, 1.53). This process of updating the pseudo-counts iterates with each subsequent token of input.

The updated pseudo-count values can be used to sample a posterior probability. For any particular parameter value with an updated value that is an increased value, the average value of the posterior probability will be greater than that of the prior probability. This is what we saw by changing the values of weights in Figure 2.2. Thus updating the pseudo-counts skews the probability distribution in favor of the parameter setting with the greater pseudo-count value. The greater the difference between the pseudo-count

values, the more the distribution is skewed, and the more likely the learner is to favor the parameter setting represented by the greater pseudo-count value. It is in this way that updated pseudo-counts affect the learner's expectations and push the learner toward favoring some parameter values over others. I will not predefine a stopping point for the learning process, but we can say that when the difference between a parameter's pseudo-count values is relatively large, then the model has learned a grammar with the parameter setting of the higher pseudo-count value. We saw in Figure 2.2 that as the preponderance of evidence accumulates for one hypothesis, the learner gets pushed strongly toward that hypothesis. Thus the dirichlet distribution is well suited to modeling a learner that has end-states in which a single value for a parameter is strongly favored to the exclusion of other values, as is typically the case with syntactic parameters that are set positively or negatively.

With the basics of the updating procedure in hand, we can now see more clearly how with priors that have weak expectations, a smaller amount of input can influence a learner to favor a particular parameter setting relatively quickly. Again, I will compare the rather extreme cases of a relatively weak prior with pseudo-counts (1, 1) and the much stronger prior of (100, 100). The lower the pseudo-count values, the weaker the prior is. Sticking with the example of the VP-headedness parameter, let us consider the effects on the probability distribution when these initial values are updated. Suppose the learner first encounters 5 tokens of input that are taken by the learner as evidence for a V-initial grammar. (This input could be ambiguous or unambiguous for V-headedness without affecting the discussion below.) The updated pseudo-counts for the different priors would be (6, 1) and (105, 100) respectively. The updated probability distributions are given below, for the updated weak prior in Figure 2.5, and for the updated strong prior in Figure 2.6.

Figure 2.5 Dirichlet probability density function, parameters = [6, 1]

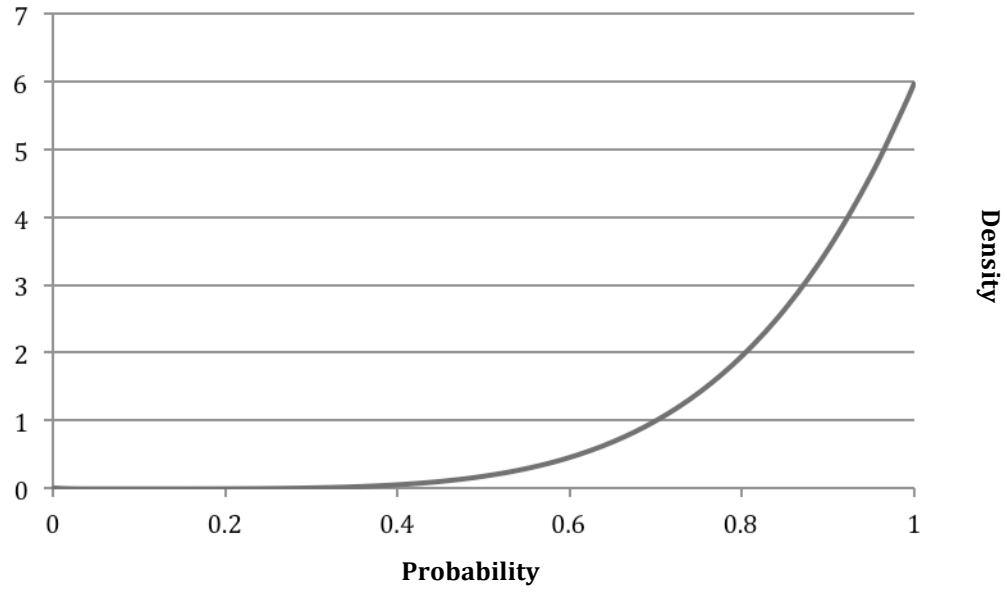
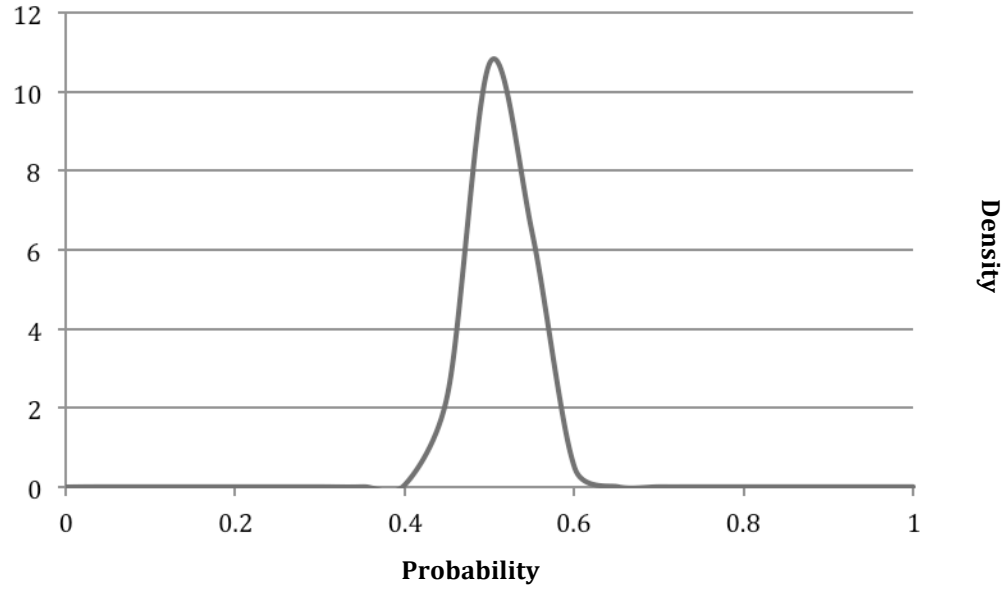


Figure 2.6 Dirichlet probability density function, parameters = [105, 100]



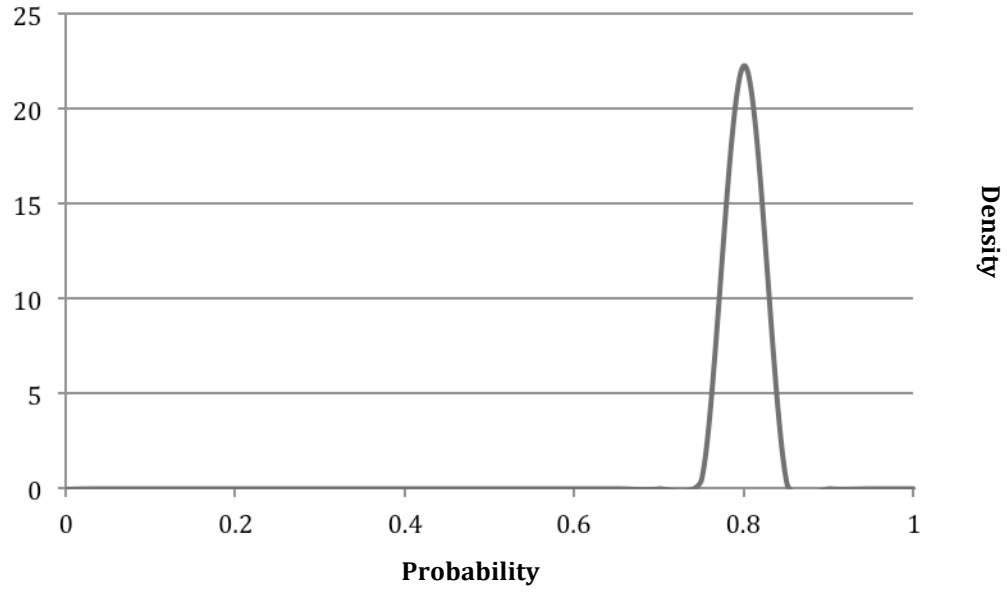
These probability distributions can be compared with those of the priors in Figures 2.3 and 2.4. A distribution with a negative skew favors the hypothesis of the first pseudo-count value (here V-initial), where a positive skew favors that of the second pseudo-count value (V-final). What we see is that updating the weak priors in Figure 2.5 strongly skews the distribution in favor of V-initial (which has the higher pseudo-count value). This means that V-initial is much more likely to have a higher probability associated with it, which in turn means V-initial is more likely to be sampled and further reinforced. In short, updating the weak priors has now strongly pushed the learner toward learning a V-

initial grammar. This dramatic push is possible with a learner that is initially relatively uncertain about what the adult grammar looks like with respect to VP-headedness. This push is also what we would expect of a learner that is less tentative: a small corpus of input (here 5 tokens) has had a large impact on this learner's grammar. In contrast, the distribution for the updated strong priors in Figure 2.6 is hardly skewed in comparison to Figure 2.5. The distribution in Figure 2.6 has a slightly longer tail on the left than the non-updated distribution in Figure 2.4, although the updated distribution in Figure 2.6 is still strongly peaked around generating a probability of 0.5 for both parameter values. A learner with such strong priors simply has not seen enough evidence to put much stock in favoring a V-initial analysis of the input. Such a learner has a much stronger degree of certainty early on that both parameter values are attested in the adult grammar at a roughly 50-50 rate. Consequently, this learner will not be strongly pushed in favor of either hypothesis after a small amount of input, and is in general less likely (in the short term) to set or mis-set a parameter on the basis of future ambiguous evidence, which supports both competing hypotheses. Thus adopting equal priors with low pseudo-count values allows for modeling a less tentative learner, as desired, as well as one that learns more quickly from a smaller corpus of input, which is helpful for expediently providing a proof-of-concept illustration of the model.

There is another way that smaller pseudo-count values can be used for the modeling aims here. For reasons similar to preferring priors with weak expectations, putting some limit on the absolute value of the pseudo-count totals will also be helpful in modeling a less tentative learner in a quick and expedient way. To implement this, at various points in the learning process I will normalize the sums of the pseudo-counts to lower values, such as 5 or 10. The process of normalizing pseudo-counts is the final component in the basic mechanics of how the model works. The discussion below focuses on an example of how this process works.

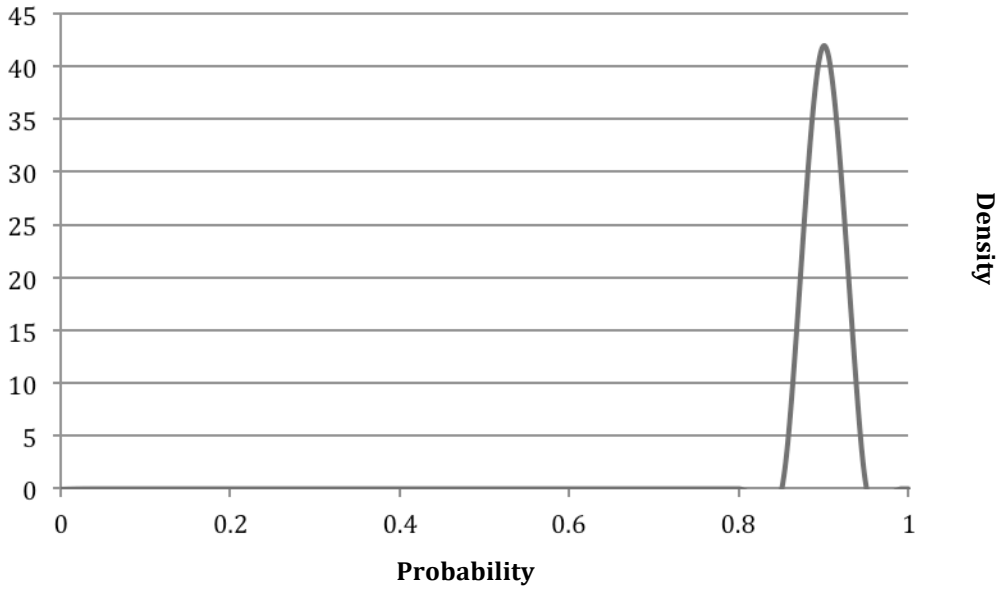
The following provides a schematic illustration of how normalizing the pseudo-count values will be helpful. I begin with a scenario in which the pseudo-count totals are not normalized. Suppose that after encountering 500 tokens of input from a V-initial language, the learner has pseudo-count values of (400, 100) in favor or V-initial. This scenario is possible in theory with a target grammar that is either V-initial or V-final. Let us first consider the case where the target is V-initial. With updated pseudo-counts of (400, 100), the learner is clearly on the way to learning a V-initial grammar. However, there has been sufficient ambiguous evidence throughout the course of learning to give a fair amount of support to the V-final hypothesis. Figure 2.7 illustrates the probability distribution with these updated pseudo-count values.

Figure 2.7 Dirichlet probability density function, parameters = [400, 100]



The skew in favor of V-initial is easy to see in Figure 2.7. However, the probability mass is tightly peaked around 0.8. This means that when generating a probability for the V-initial parameter setting, this probability is unlikely to deviate much from 0.8. Such a probability certainly favors V-initial, but as this probability is not likely to get much closer to 1 (given the current weights), the learner will not be able to readily reject the V-final hypothesis. Indeed, it would take hundreds more tokens of input to push the peak considerably toward 1. In the unrealistic situation that the learner never subsequently increased the pseudo-count value for V-final, it would take 500 additional tokens of input to push the peak to 0.9, as illustrated in Figure 2.8.

Figure 2.8 Dirichlet probability density function, parameters = [900, 100]



I note that the learner's development reflected in the change from Figure 2.7 to Figure 2.8 is not representative of the case studies discussed in this work. The ambiguous nature of the input in these case studies means that we would expect much less of a skew than shown in Figure 2.8 after an additional 500 tokens of input. The ambiguous input would prolong the learning process by making it even more difficult to reject the V-final hypothesis.

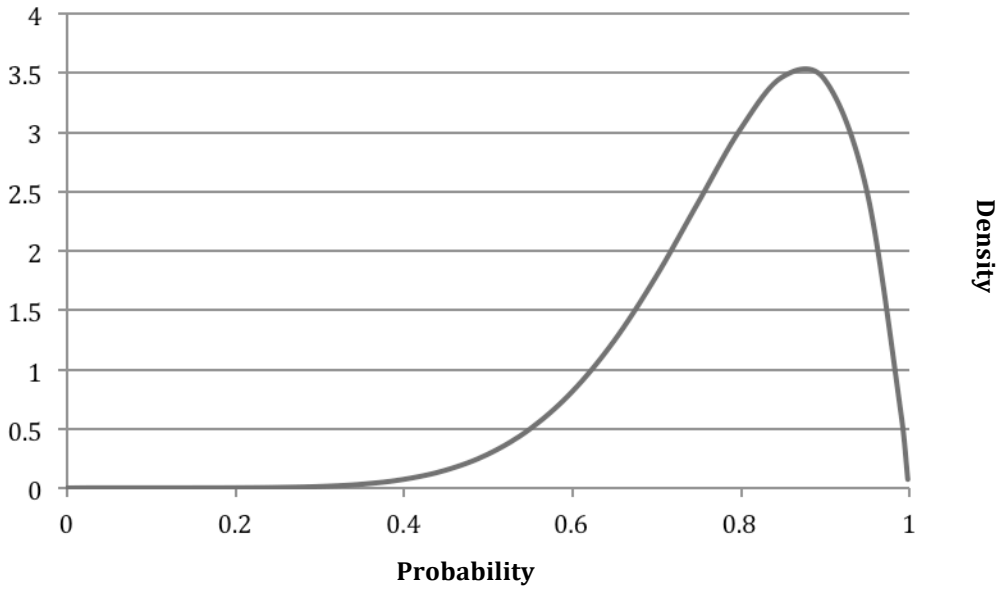
The distributions in Figures 2.7 and 2.8 represent a relatively tentative learner. The learner slowly moves toward a V-initial grammar, and at each point in time, the probability generated for a V-initial grammar deviates relatively little. Throughout much of the early stages of learning, then, a V-initial grammar will not be sampled with a very high probability.

A similar logic holds in the case where the target grammar is V-final. Suppose we want to model a learner that first mis-sets the parameter to V-initial before recovering and learning a target setting of V-final. In Chapter 3, I show how a comparable situation can arise in child Swiss German given ambiguous input and parameter interaction. As we have seen in Figures 2.7 and 2.8, the learner will certainly favor the non-target grammar, but the learner will be reluctant to fully adopt the non-target parameter setting.

To expedite the learning process, it will be helpful to periodically normalize the pseudo-count totals to lower values. This normalization will more quickly increase the likelihood that some parameter is sampled at a high frequency, which results in a less tentative learner. To illustrate this, consider the scenario from Figure 2.7 above but with the pseudo-count values normalized so that they sum to some relatively small total, such as 10. The proportion of the pseudo-count values remains the same, but now the values are (8, 2) for V-initial and V-final respectively. The probability distribution of these

normalized values is given graphically in Figure 2.9. Again, let us first consider the case where the target grammar is V-initial.

Figure 2.9 Dirichlet probability density function, parameters = [8, 2]



If we compare Figures 2.7 and 2.9, we see that in Figure 2.9 higher probabilities are represented by a substantially larger area under the curve. This means that it is much more likely for a high probability to be generated for V-initial, which will greatly increase the sampling frequency of V-initial. With a greater sampling frequency, V-initial is more likely to be reinforced. If the process of normalization is repeated throughout the learning process, V-initial will be more and more likely to be reinforced, which allows for more rapid learning of a V-initial grammar. Even with abundant ambiguous evidence regarding VP-headedness, the more extreme skew of the distribution allows the model to learn a V-initial grammar with relative ease. Similarly, learning a non-target grammar is less of a challenge with normalized pseudo-counts. If we want to model learners who at first incorrectly adopt a V-initial grammar, V-initial is more likely to be sampled given the distribution in Figure 2.9 (with normalized pseudo-count values) than the one in Figure 2.7. In sum, normalizing the pseudo-count values allows for modeling a less tentative learner. This learner is more likely to quickly set or mis-set a parameter in the face of ambiguous evidence. I will thus adopt this normalization approach when implementing the model for the proof-of-concept illustrations of the case studies in subsequent chapters.

I have now introduced the basic outline of how the model will be implemented. More specific details concerning the simulations for each case study can be found in the results sections of Chapters 3-5.

3. Comparison with other models

In this section I introduce three other models for language learning and compare them to the model proposed in this work. After briefly describing the models, I consider how they fare with the phenomena discussed in the Chapter 1. Can these models reliably learn the grammar of a subset language or some other parameter setting given only ambiguous evidence? Are they able to model learner errors and variability? Although some of the models do well with some of these learning tasks, none is able to address all the learning challenges that are the focus of this work. A strength of the current model, then, is that it provides a unified approach to all these phenomena.

The models that are used for comparison by no means exhaust the different kinds of learning models that have been proposed in the literature. Rather, they have been chosen as being representative of three distinct approaches to learning. The purpose of this comparison, then, is to identify relative strengths and weaknesses of these different approaches (as exemplified by specific models) and to contrast them to the current model.

The three alternative models discussed below are Sakas and Fodor (2001), Gibson and Wexler (1994), and Yang (2002). The model in Sakas and Fodor (2001) is what Sakas and Fodor (2012) describe as a deterministic model. It attempts to incrementally arrive at the target grammar by always setting parameters to their target values. It does this by only learning from unambiguous evidence. In contrast, the model in Gibson and Wexler can use ambiguous evidence to set parameters, but it does so in a rather naïve way. In their model, any grammar that is compatible with ambiguous input is just as good as any other. There is thus no notion of a grammar of best fit when it comes to ambiguous evidence. Finally, Yang's (2002) model is also a probabilistic model that has a strong similarity to how the current model learns from ambiguous evidence. The research program initiated here can be seen as building on Yang's work in showing how a similar model can be applied to a new range of empirical phenomena. However, it also goes beyond Yang's model in accounting for learning puzzles where Yang's model falls short.

As discussed in the introduction to this chapter, these three models are all discriminative models. They all evaluate whether a token of input is compatible with a particular grammar. As I have just described, these models differ as to what kind of input is used to evaluate compatibility. For Sakas and Fodor (2001), the input must be unambiguous for a particular parameter value. In the case of the subset learning scenario there simply is no unambiguous evidence to learn the grammar of the subset language, and the learner is not even in a position to evaluate whether the grammars of the subset and superset languages are input-compatible. More generally, though, these models are all discriminative at their core and thus contrast with the generative model I have presented in being able to systematically learn the grammar of subset language. For a discriminative learner the grammars of both the subset and superset languages are equally compatible with all the input available to the learner. In the generative model, it is the grammar of the subset language that is a better fit to the corpus of input.

In sum, the three alternative models represent approaches to learning that have varying degrees to which ambiguous evidence plays an important role. This ranges from none at all for Sakas and Fodor, to a somewhat incidental role for Gibson and Wexler, to a fundamental role for Yang. My remarks in this section will be of a more general nature. At the end of each case study in Chapters 3-5, I will consider all of the models again in more detail with the particulars of each case study in mind.

3.1 Sakas and Fodor (2001): The Structural Triggers Learner

The goal of Sakas and Fodor's Structural Triggers Learner (STL) is to move incrementally closer and closer to the target grammar without mis-setting any parameters. The learner begins in an initial state without any values being adopted for any parameter setting. Every time the learner sets a parameter, it does so after having encountered unambiguous evidence for that parameter setting. Thus the (partial) grammars that the learner adopts throughout the course of learning are characterized by a monotonic increase in the set of parameters that have been set correctly.

In this model, the parser builds a parse tree of the input. At a given point involving a parametric choice, the parser is able to recognize whether this choice is underdetermined given the input data or whether a parse is possible with only one possible parameter setting given the input. If only one choice is compatible with the input, this constitutes unambiguous input for that parameter, and the parameter is set accordingly. This is the discriminative component of the model: if the input is compatible under a parse with only one of a particular parameter's values, then that value must be adopted by the learner. However, if there is ambiguity for a given choice point, the parser then reports this ambiguity to the learning mechanism, and the learning mechanism will not use this ambiguous input to set that parameter. In other words, the learning mechanism will wait until unambiguous input for any parameter *P* occurs in the input before setting *P*.

A fundamental challenge for this model, then, is trying to set parameters for cases in which there is no (or vanishingly rare) unambiguous evidence. A basic example of this is the subset case. Verb movement and head-complement directionality parameters in Korean provide what appears to be another example of this.[10] Because the model must wait around for unambiguous input before setting the relevant parameters, these parameters can end up never being set. These grammars then become unlearnable. I note that one obvious way of augmenting the STL is to provide it with defaults for parameters.

[10] Sakas and Fodor (2001: p. 227, n. 6) appear to assume the learner makes use of the Subset Principle. That is, assuming the learner can compute what is the grammar of the subset language and the grammar of the superset language, the learner will at first adopt the grammar of the superset language until some unambiguous evidence for the grammar of the superset language is encountered. This would help the learner acquire the grammar of the subset language if that is the target. Nevertheless, the case of Korean, which does instantiate the subset learning scenario, presents the learner with a dearth of unambiguous evidence and does not allow the learner to invoke the Subset Principle as a way out of this bind.

As Sakas and Fodor (2012) discuss, these defaults represent 'guesses' (that are consistently of the same value for the same parameter across languages and speakers) that can be used until unambiguous evidence is encountered. If unambiguous evidence is never encountered, then the learner will maintain that guess through to the end-state of learning. In the subset case, for example, the default could be the parameter value that leads to the grammar of the subset language (but see note 12). Introducing defaults to the learner is not without its own complications (cf. Sakas and Fodor 2012 on this point), and I will discuss the ramifications of this when we have the details of the case studies in hand in Chapters 3-5, as well as in the conclusion in Chapter 6.

Moreover, the STL cannot model learners' errors or variability. The reason for this is again because of its exclusive reliance on unambiguous evidence. With only unambiguous evidence to learn from, the model cannot mis-set a parameter. But without mis-setting a parameter, it is not clear how the model would account for learner errors with a non-target parameter value. Again, the use of defaults might capture some errors. The STL would then become an error-driven learner. Errors could occur if the learner's grammar contains some default values that are non-target. These non-target default values would eventually get replaced with the target values when the learner is exposed to the relevant unambiguous input. But before the default values get replaced, the learner would have a grammar that would allow errors according to the default values. I discuss this in more detail in later chapters.

Further, the variability in verb raising I consider with Korean in Chapter 4 is possible precisely because it appears there is no (or vanishingly rare) unambiguous evidence to guide the population to a uniform analysis of the input. If the input for the whole population contained some unambiguous evidence either for or against verb raising, then that is the parameter setting that the entire population would adopt. Variability also provides us with a clear example of how defaults will not help this kind of deterministic learner. Lacking (sufficient) unambiguous evidence to the contrary, the population of learners will uniformly stick to the default guess, and variability does not arise.

In its basic form, then, the STL has a number of limitations because it essentially filters out all ambiguous evidence. As was discussed in Chapter 1, the most promising approach to address these limitations is for the model to use defaults for parameters. Without unambiguous evidence for a particular parameter, the learner could rely on a default value for that parameter. Nevertheless, the use of defaults still does not allow such a learner to model variability in language learning when this variability is driven by ambiguous evidence. Further, we expect errors to reflect only the default value, as the STL cannot mis-set a parameter. The contrasting error patterns mentioned in Chapter 1, then, are a problem for an approach that relies on defaults. In contrasting error patterns, we see learner errors that reflect multiple values of the same parameter, and thus cannot be due to a single default. I discuss defaults and the details of contrasting error patterns in

Chapter 6.[11]

3.2 Gibson and Wexler (1994): The Triggering Learning Algorithm

The Triggering Learning Algorithm (TLA) in Gibson and Wexler (1994) allows the learner to set parameters on the basis of ambiguous evidence, but it does so in an uninformed way. In the initial state of the learning process, the learner starts with some (possibly random) set of values for all the parameters. The TLA is discriminative, and in particular it is error-driven: when the learner encounters some input that is incompatible with the current grammar, the learner must try to change its grammar to find a compatible grammar. This involves changing a parameter setting. If the input is ambiguous, though, all input-compatible grammars that the learner could move to are equally likely to become the new grammar. The TLA does not make use of parameter interaction to learn a grammar that is 'most compatible' among the compatible grammars. As such, the model cannot use ambiguous input to systematically learn some kinds of compatible grammars over others.

The way that the TLA changes a parameter setting is restricted by two constraints: the Single Value Constraint and the Greediness Constraint. According to the Single Value Constraint, when the learner encounters new input that is incompatible with the current grammar, the model will attempt to adopt a new, compatible grammar by changing at most one (possibly randomly selected) parameter setting. If this new, minimally different grammar is compatible with the input, then it will be adopted by the learner. The Greediness Constraint forces the learner to only adopt a new grammar if it input-compatible. Thus if the grammar resulting from changing a parameter value is not

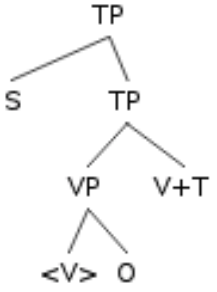
[11] As another possibility for augmenting a deterministic learner such as the STL, Fodor (1998: 26) suggests that prior to setting a parameter via unambiguous evidence, a parameter remains "unset", and either value can be used "freely" in speaker productions. This suggests that when producing utterances with unset parameters, the learner is sampling a value from an unbiased probability distribution. I will not consider this possibility any further for the following two reasons. First, I have claimed that in some cases, defaults are necessary for a deterministic learner because of the absence of unambiguous evidence. It is not clear how the learner could wait around to set an "unset" parameter while also availing itself of a default. As the learner can reach a final state in which all parameters are categorically set, even when there is an absence of unambiguous evidence, we can conclude that the STL needs default values and is not agnostic about values in the initial state. Second, although random sampling of an "unset" parameter value each time the learner produces an utterance could of course result in *some* non-target utterances with respect to a particular parameter value, we do not expect near-ceiling error rates in a large sample of utterances for that parameter value. This is because nothing would prevent the learner from sampling both parameter values to produce utterances that reflect a range of different grammars. In Chapter 3, I discuss how we see near-ceiling error rates in embedded verb placement in the productions of Swiss German children during an early developmental stage. In my analysis, this can be attributed to the children adopting a grammar with a non-target parameter value. Under the alternative of randomly sampling parameter values that Fodor (1998) proposes, these high error rates are unaccounted for. At some point during the developmental stage in question, we would expect to see, at some non-vanishingly rare rate, utterances of the relevant type without the error as a result of sampling the target (or a target-like) grammar, when in fact they are almost never attested. In contrast, the STL could account for these high error rates if the children's grammar involves a non-target default value for a parameter that has not yet been set.

compatible with the input, then the learner will maintain the old (non-target) grammar, and must wait for further input that will force the learner to adopt a new grammar. The intention of the Greediness Constraint, then, is for the learner to always be adopting new grammars that are closer to the target grammar. Crucially, the TLA only learns if the current grammar is incompatible with the input. If a particular parameter has more than one value that is compatible with all the input, the TLA can use either one and will never need to change that value. Such is the case when learning the grammar of a subset language or head-finality in Korean. In contrast to a probabilistic model that can learn the grammar of best fit, the TLA cannot systematically learn the grammar of the subset language instead of the grammar of the superset language, say, or learn that a verb-final language such as Korean is consistently head-final.

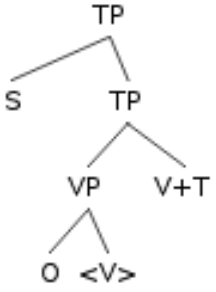
Just as with the STL, adopting defaults could help the TLA learn the grammar of the subset language if that grammar is the target. The learner would then never move to the grammar of the superset language because the grammar of the subset language is compatible with all the input (but see note 12). Adopting default values, though, does not resolve all potential complications stemming from ambiguous input, as is discussed below.

The TLA fares somewhat better with modeling learner errors and variability. We can in fact see how both are possible in the TLA with the same example. Recall from (6), repeated below, that there are various grammars compatible with SOV input in Korean:

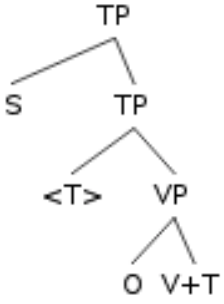
(6) a. *Verb raising Korean*



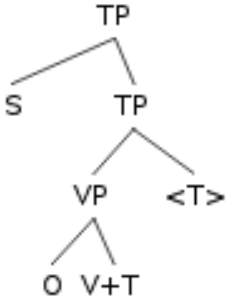
a'. *Verb raising Korean*



b. *Verb in-situ Korean*



b'. *Verb in-situ Korean*



Suppose two learners both start with a [–V-to-T, V-init, T-fin] grammar. This is not one of the grammars in (6) and is not compatible with SOV input. Thus this grammar could

be used by the model to account for any non-target SVO utterances that the learners might produce. According to the TLA, upon hearing SOV input, both learners would attempt to change a parameter setting. One learner could change V-initial to V-final, resulting in grammar (6b'). The other learner could change [–V-to-T] to [+V-to-T], and would then adopt grammar (6a). For the sake of illustration, let us then suppose that no further input would force these two learners to change the parameter settings in the new grammars they have adopted. (In Chapter 4, I will claim that these kinds of grammars are indeed compatible with essentially all Korean input and thus are viable grammars for adults.)

With this simple illustration we can see how the TLA has no difficulty in allowing for variability across speakers: some speakers will have verb raising while others do not; and some speakers will have some head-initial phrases while others have those same phrases as head-final. In fact, we see more variability with the TLA than we do with the probabilistic model I have proposed: whereas both models predict variability for verb raising, the probabilistic model will push all learners toward head-finality for all the relevant phrases. The null hypothesis is that for any given parameter, Korean speakers have the same parameter setting. Both the TLA and the learning model proposed here provide evidence against the null hypothesis. What currently distinguishes them is the additional evidence that can be brought to bear on the question of variability. As discussed in Chapter 4, there is experimental evidence from Han et al. (2007) supporting the claim of variability in verb movement across learners of Korean. There is no additional evidence suggesting the TLA's greater range of variability is attested in Korean. An empirical question that can be more carefully investigated is whether there is any experimental evidence for this additional variability. Assuming this additional variability is not attested, then how might this range of variability be decreased?

The TLA could make use of default values for parameters to decrease the range of variability. If all learners started with head-final values for the head-directionality parameters (7), there is no input that would cause a change in the values for these parameters.

(7) *Partial set of default values*
[T-fin, V-fin]

As with the probabilistic model, then, we would see uniformity across speakers in adopting a consistently head-final grammar. However, lack of variability in head-complement direction would also result in a lack of variability for verb raising. In the simple example in (6), both uniformly head-final grammars in (6a') and (6b'), regardless of whether there is verb raising, are compatible with the input. Again, there is nothing in the input to force the learner to adopt a grammar with a different value for verb raising. In Chapter 4, I discuss how this example extends naturally to Korean, and how adopting default values for all parameters precludes the TLA from capturing variability for verb

raising in Korean. By tightening up the predictions of the model with the defaults in (7), the TLA ends up in a similar position to Sakas and Fodor's (2001) STL when it is augmented with default values: both models have limitations in accounting for variability when presented with ambiguous evidence. With defaults, both models are error-driven in that they will adopt a non-default value only when presented with evidence that is incompatible with the default grammar. In cases such as Korean, where I claim that the learner receives essentially only ambiguous evidence that is compatible with a default grammar incorporating the values in (7), such a default grammar prevents these models from learning anything further from ambiguous input.

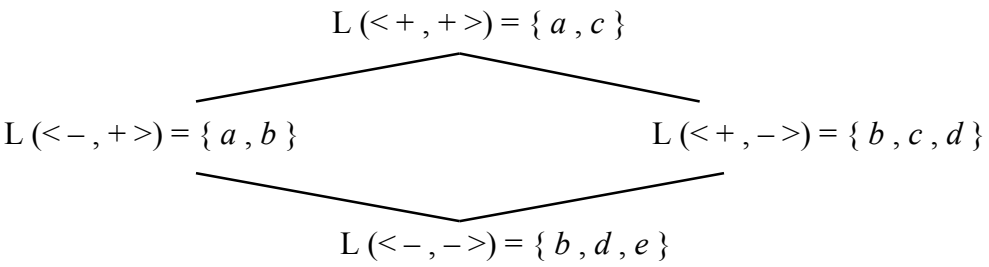
In sum, we have seen mixed results in how the TLA accounts for variability. It can account for variability, but it predicts a greater range of variability than is predicted by my learning model. This greater range of variability is not currently supported with experimental results. This is not a decisive argument against the TLA, but the balance of evidence is currently in favor of the model I have proposed. In attempting to constrain the variability with a particular set of defaults, though, the TLA is no longer able to account for any variability. An open empirical question, then, is to what extent the variability predicted by the TLA can be supported by experimental findings.

The TLA can also model variable errors that a single speaker produces during the course of learning. In the case of Swiss German discussed in Chapter 3, children pass through a stage in which their productions of embedded clauses are characterized by variable placement of the finite verb. Sometimes it is in the target, clause-final position, and sometimes it is in the non-target, non-clause-final position. As Frank and Kapur (1996) illustrate, it is possible for the TLA to move back and forth between the same two grammars. Frank and Kapur (1996: 633) provide the formal example in (8).[12] The set of languages L has four members, and the sets of letters correspond to the different types of utterances that are possible in that language. These utterances correspond to the different types of input that a learner would receive for each of those languages. There are two

[12] Frank and Kapur assume that there are no subset/superset languages, and indeed none are present in the example in (8). As was discussed in Chapter 1, I do assume that there are subset/superset languages. Nevertheless, there is good reason to exclude them from an example such as (8). This relates to a shortcoming of the TLA or STL in being able to successfully converge on the grammar of a subset language. I have suggested that the TLA and STL could successfully acquire the grammar of a subset language by starting with that grammar in the initial state. This will not always help in cases where there are grammars of multiple subset languages. Consider a modification to (8) in which L (< + , + >) = { a }, and L (< – , – >) = { b }. There are now two subset languages, but which one should be the initial state? (There are also two superset languages, but that is incidental: only { a , b } is necessary to illustrate the following point.) Regardless of which subset language is the initial state, in the case of the TLA, the learner would never be able to move from an initial state of one subset language to a target end-state of the other subset language without getting stuck in the local maximum of a superset language. In the case of the STL, there is insufficient unambiguous evidence to choose between the grammars of the < – , + > superset language and the other subset language, and the learner would remain stuck in the non-target initial state. Again, the model I have proposed could use implicit negative evidence to address this challenge along the lines discussed in this chapter. This scenario thus constitutes another challenge for the TLA and STL and further complicates the practicality of the use of defaults or the Subset Principle with the TLA and STL. For reasons of simplicity, though, I have abstracted away from this point in the main text.

binary parameters; thus the four languages represent the full set of logically possible grammars. The parameter values of each language are represented schematically with an ordered pair of pluses and minuses. Each grammar is connected by a line to another grammar that is minimally different by one parameter value (i.e. the difference between the grammars satisfies the Single Value Constraint). Frank and Kapur focus on the case where the target language is < – , + >, and the initial state of the learner is < + , – >.

(8)



The only kinds of evidence in the target language are *a* and *b*. If the learner hears *b*, no change in grammar is necessary. It is only when the learner hears *a*, which is not possible in the initial grammar, that the learner must try to adopt some new parameter value. The only new grammar that satisfies the Single Value Constraint and the Greediness Constraint is < + , + > , which is also compatible with *a*. If the learner continues to hear *a*, again no change is necessary. But if the learner then hears *b* again, there are two viable grammars that the learner could adopt, both of which satisfy the constraints on learning. One is the target < – , + > grammar; the other is the original < + , – > grammar. Both grammars are equally good choices: the learner, of course, does not know what the target grammar is, and the learner has also not been keeping track of which grammars it has adopted in the past. If the learner chooses the target grammar, then no further changes in parameter settings will be made. If the learner returns to the grammar of the initial state, then the learner will eventually return to the < + , + > grammar upon hearing *a*, and the back and forth alternation between grammars will repeat until the learner finally chooses the target grammar.

This back and forth alternation is a way that the TLA can model a learner's variable errors. Under the non-target < + , – > , the learner can produce utterances of the form *d*, but this option disappears when the learner adopts the < + , + >. The possibility of producing *d* reappears, though, when the learner returns to the original < + , – > grammar. Crucially, this kind of variability is only possible in the TLA when the learner moves back and forth between multiple non-target grammars. This is because the learner will never change any parameter settings once reaching the target grammar; no input the learner receives is ever incompatible with the target grammar's parameter settings. In cases of variable errors, then, the TLA predicts that when the learner adopts the < + , + > grammar, utterances of the form *c* are still possible. We might expect the learner to alternate between producing errors of the form *d* or not, while still producing errors of the form *c* until finally reaching the target grammar. The case of Swiss German does not

appear to support this prediction. In the data on the Swiss German children reported by Schönenberger (2001), the absence of verb placement errors does not appear to always correlate with some other con-current error. Thus, although the TLA is capable in principle of accounting for a learner's variable errors, it does so in a rather inflexible way that does not model actual learners' development. In the probabilistic model I have proposed, it is possible to have heavy weights favoring all but one of the target parameter values. Errors are then expected to be due only to the final unset parameter. The weights associated with the final parameter can distributed across this parameter's values such that sometimes these errors are possible, and other times they are not. In this way it is possible to model a learner that alternates between the target and a non-target grammar.

In sum, although the TLA does not filter out ambiguous evidence, it cannot learn from this evidence in a highly informed way. In a learning scenario such as the subset case, this means that the TLA cannot reliably learn the grammar of the subset language without some additional component such as default. Further, although the TLA is not incompatible with modeling errors or variability, the learning outcomes it predicts concerning these phenomena are different from those under the probabilistic learner I have proposed, and these outcomes do not always correspond with acquisition data on children's development.

3.3 Yang (2002): The Naïve Parameter Learner

Yang's (2002) Naïve Parameter Learner (NPL) is a probabilistic model that works much the same way as the simplified version of the model presented in Section 2. There is a crucial difference between the models, though, and this emerges when we consider the full version of the model from Section 2, which learns from implicit negative evidence. Like Sakas and Fodor's (2001) STL and Gibson and Wexler's (1994) TLA, but unlike the model I have proposed, Yang's NPL struggles to consistently learn the grammar of a subset language when that is the target. Further, in this section I also discuss that even if we were just to compare the simplified version of the model I have proposed with Yang's NPL, although the NPL is capable of learning via parameter interaction, Yang stops short of showing how parameter interaction plays a role in the learning process. In a sense, the simplified implementation of the model I have introduced fleshes out Yang's proposal and extends the range of its possible applications. It is to be understood, though, that the research program here goes beyond looking at questions that Yang does not investigate. As I have discussed in this chapter, the simplified version of the model is embedded in a more complex generative model. The expectation is that this generative model can be used with comparable results for all the case studies here that use the simplified version.

In the NPL, each parameter has different probabilities associated with its different values. Similar to the simplified model in Section 2.1, when the NPL is presented with new input, a value for each parameter is sampled according to these parameters. If the resulting grammar composed of all these values is compatible with the input, then all

those values are 'rewarded' (i.e. their probabilities are increased), and all the values not sampled are 'punished' (i.e. their values are decreased). If the resulting grammar is instead incompatible with the input, then the values that were sampled are 'punished', and the values not sampled are rewarded. We can see that this simple evaluation of whether a particular grammar is input-compatible distinguishes the NPL as a discriminative learner. The NPL is called 'naïve' because of how it responds to ambiguous evidence: it will sometimes reward non-target parameter values that are compatible with the input. Despite this naïveté, Yang discusses that in the long run the NPL can learn the target grammar if there is sufficient unambiguous evidence that can reward some parameter values while punishing others.

I note that a further difference between the NPL and both versions of the model I propose concerns the belief-states or expectations of the learner. For Yang the weights associated with different hypotheses are simple probabilities (e.g. .5), as in the exposition in Section 2.1. This contrasts with the refined implementation of weights I introduced in Section 2.4, in which weights are pseudo-count values of the dirichlet distribution. The difference can be characterized as follows. At any point in learning with the NPL, the learner has 100% certainty as to what the actual probability of a particular parameter setting is in the adult grammar. For example, if the current weight for V-initial is .6, then learner is 100% certain that V-initial is attested in the adult grammar with a likelihood of 60%. This contrasts with the belief-states of the model I have proposed. In the model in Section 2, at any given point the learner has varying degrees of certainty (based on the strength of the evidence) as to what the actual probability is. For example, at a given point in time, a learner that has only moderately strong expectations about the adult grammar might only be 80% certain that V-initial is attested in the adult grammar with a likelihood of 60%. In Section 2.4 I discussed how the dirichlet distribution is able to model varying strengths of a learners expectations, and we saw how this is a psychologically more plausible approach to modeling learning.

Returning now to Yang's NPL, a representative example of how the NPL learns is presented here. Yang considers the case of learning a V2 language such as Dutch. As a simplified set of input, the learner will encounter the types of matrix clauses in (9a), where X is some initial constituent other than the subject, object, or the finite verb. The target grammar for Dutch represents one hypothesis, but there are other grammars the learner must rule out, and these other hypotheses are represented by other attested languages in (9b-e). Each language is accompanied by a small set of characteristic input types that a learner of those languages would encounter.

(9) a. Dutch: SVO, XVSO, OVS
b. Hebrew: SVO, XVSO
c. English: SVO, XSVO
d. Irish: VSO, XVSO
e. Hixkaryana: OVS, XOVS

(Yang 2002: 35)

The set of input for a Dutch learner overlaps with the sets of input that are compatible with every other grammar under consideration in (9). Thus Yang observes that every token of input that a Dutch learner receives is ambiguous: it is compatible with multiple grammars. The grammar of Dutch, though, clearly has the greatest likelihood of being compatible with the input: all other grammars can be punished in some way, but the Dutch grammar never will be. For example, if the learner encounters OVS input and samples parameter values that yield a Hebrew, English, or Irish grammar, those grammars are not compatible with the input and the probabilities associated with their parameter values will be decreased. When the input types in (9a) are assigned relative frequencies based on frequencies of comparable utterances in a corpus of child directed Dutch, Yang demonstrates with a simulation how there is a sufficient amount of each input for the NPL to learn the Dutch grammar.

The NPL goes beyond Gibson and Wexler's TLA in actively learning from each token of ambiguous input. In Section 3.2 we saw that the TLA could use ambiguous to move from one grammar to another. However, subsequent occurrences of ambiguous input did not necessarily have any effect on the learner. So long as the new ambiguous input was compatible with the most recent grammar, the learner would not change any parameter settings. In contrast, Yang's NPL can use each token of ambiguous input to either reinforce or punish the parameter values of whichever grammar is sampled.

Ambiguous evidence thus plays a central role in how the NPL learns. Indeed, the general contours of the NPL follow those of the simplified model from Section 2, in which parameter values that are reinforced are only those that are sampled and compatible with the input. Thus the NPL can also learn effectively via parameter interaction for the reasons that have already been discussed. This means that the NPL is capable of modeling the kinds of learner errors and variability we see in Korean and Swiss German.

What is important to emphasize here is that Yang's work does not actually explore the application of parameter interaction to modeling learner errors and variability.[13] The example in (9) takes a very broad view of learning different languages and does not look

[13] See Wexler (2011) for a related point about how Yang does not present any empirical examples that investigate parameter interaction. Yang (2002: 36-39) acknowledges the possibility that parameters might interact. Indeed Yang (2002: 51) claims to have run a simulation with 10 interacting parameters that can converge on a particular target grammar. Unfortunately, as Wexler (2011) notes, Yang does not provide any information about the nature of these parameters (e.g. whether these parameters correspond to those found in natural languages, and to what extent the parameters interact with each other, etc.). Thus it is difficult to evaluate the results of Yang's simulation. Moreover, Yang's discussion of parameter interaction overlooks the general application of parameter interaction to understanding learners' development. For Yang, such interaction is presented as a potential problem that might prevent the learner from acquiring the target grammar, although the empirical examples he illustrates do not actually involve any parameter interaction. Far from being a problem, a contribution of the current study is to show that parameter interaction in Korean and Swiss German helps to shed light on modeling learners' development. Even when parameter interaction does push the learner away from the target grammar, as in the case of Swiss German, this has the welcome effect of capturing children's errors. Further, such parameter interaction is not sufficient to prevent the model from learning the adult grammar for Swiss German.

at the subtle interplay between all the different parameters values that give rise to the languages represented in (9a-e). We can contrast the example of Dutch in (9) with the example of Korean in (6). In Yang's example (9), although every type of input is compatible with multiple grammars, only one grammar is compatible with every type of input. This observation can be sufficient to see what grammar the model will learn; digging deeper to explore parameter interaction is, in a sense, superfluous. The Korean example in (6) presents a different learning situation. In (6) we saw that there are multiple grammars that are compatible with all the input. In Chapter 4, I show how this generalization extends to a much larger corpus. In such a learning scenario, the role of parameter interaction takes center stage.

In other examples (e.g. Yang 2002: 38, 43-44, 103-104), Yang does look at individual parameters, but these parameters are considered either in isolation, or they do not interact with any of the other parameters under consideration. Thus, the empirical phenomena of Korean and Swiss German concerning variability and errors fall outside the scope of Yang's work, though they can be captured by the NPL. In contrast, one focus of the current study is to investigate these very interactions, thereby extending Yang's basic approach to a more complex domain.

There is, moreover, a more fundamental difference between the NPL and the full version of the model I have proposed. When we set aside the possibility of parameter interaction, the NPL struggles to consistently learn the grammar of a subset language when that is the target. As all the evidence the learner receives is compatible with either the grammar of the subset language or the grammar of superset language, there is no principled reason why the NPL should always punish the grammar of the superset language. The situation is much like flipping a coin. Either parameter value is a viable option, and if the NPL samples the same value enough times, by chance the learner might go through a stretch where one value is sampled sufficiently more than the other such that the learner will adopt that value as a parameter setting. Indeed, as Pearl (2007) discusses when looking at one parameter in isolation, so long as the NPL is not too conservative or tentative of a learner, then when presented with only ambiguous evidence, the NPL could learn either parameter setting. In the case at hand, this means that the NPL could learn either the grammar of the subset language or the grammar of the superset language. The trouble for the NPL is that it has no way of learning from implicit negative evidence. The full version of the model I have proposed can cash out on the learner's additional expectations under the grammar of the superset language, thereby using ambiguous evidence to learn the subset grammar. [14]

[14] In response to the learning challenge posed by the subset scenario, one could imagine a variant of the NPL that selectively had very high prior probabilities for the parameter values that distinguish the grammar of the subset language from the grammar of the superset language. These biased priors would be like having defaults, but having them only for parameters that crucially characterize the grammar of the subset language. Stipulating such priors for these parameters is an unnecessary complication, given that the

The probabilistic learner I have introduced thus draws on Yang's work to present new insights regarding parameter interaction while also showing how the learning process can address the challenge presented by the subset scenario.

4. Summary

In this chapter, I have introduced a probabilistic learner that provides a unified approach to the learning challenges of ambiguous evidence by learning from that very same ambiguous evidence. I have discussed how it can learn from implicit negative evidence in order to learn the grammar of a subset language, as well as from parameter interaction. Shining a light on parameter interaction has shown how learner errors and variability can be modeled. These phenomena and learning scenarios are taken up with more detail in the case studies in the following chapters. I have also compared this model with three alternative models. These other models all represented different approaches to learning from ambiguous evidence. The models have varying degrees of success in accounting for all the learning scenarios under consideration; in contrast to the current model, none of these alternative models provides a straightforward account of the full range of learning scenarios I investigate.

---

learning model I have proposed can simply learn the grammar of the subset language, and I will not consider this variant any further.

Chapter 3

The Acquisition of Verb Movement in Swiss German: Modeling child production errors and variability

1. Introduction

In this chapter I explore an acquisition puzzle in children's productions of Swiss German (Lucernese dialect), which is based on the in-depth study in Schönenberger (2001, 2008). The puzzle is as follows. The canonical position of the finite verb in embedded clauses with an overt complementizer in Swiss German is clause-final. This is shown in (1a) with the complementizer *wenn* "when, if". However, in spontaneous speech the Swiss German children in the study produce non-target embedded clauses with a non-final finite verb as in (1b). After a period of time, the children begin to gradually produce more and more target embedded clauses, approaching adult-like performance by the end of the period of observation.

(1) a. Wenn t'Hex före-chunt…
when the-witch outside-comes
'When the witch comes outside…'

b. # Wenn chunt t'Hex före… (M: 3;10)
when comes the-witch outside

(Schönenberger 2001: 82)

The central question is why children are producing such non-target forms. This question is especially pertinent in light of Schönenberger's claim that the same children have largely error-free matrix clauses concurrent with the period of errors.

There are two aspects to this puzzle that make this an ideal case study for the kind of probabilistic model of learning I have been advancing. First, I will claim that the children have mis-learned some parameter setting before gradually resetting it. In this chapter I show that both parameter mis-setting and resetting can be accounted for under a probabilistic model that learns from ambiguous evidence by means of competing hypotheses that interact with each other. This interaction is parameter interaction, which was introduced in Chapter 1. Second is the variability of the children's errors, which can be accounted for under a probabilistic model with competing hypotheses.

As regards parameter mis-setting and resetting, the explanation for this developmental trajectory lies in how the model learns from ambiguous evidence via parameter interaction. According to the analysis pursued here, the Swiss German input to the learner is highly ambiguous. Nevertheless, parameter interaction can drive the model to learn some non-target parameter setting from ambiguous input. Thus, among a set of

competing hypotheses, the learner temporarily settles on a non-target hypothesis that is nevertheless a good fit to the input. Over time, as more parameters are set correctly, a new grammar of best fit emerges from among the competing hypotheses, and the learner is able to recover and set the parameter correctly. Ambiguous evidence, then, plays a crucial role what can be call a kind of mis-learning, which we see initially in the developmental course of the children.

In particular, it is the ambiguity that is found in matrix clauses that contributes to parameter mis-setting. An intuitive approach to the acquisition puzzle is to relate the position of finite verbs in matrix clauses with the embedded clause errors. Matrix clauses are characterized by a V2 word order in which the finite verb is in second position, which is frequently non-final. As will be discussed, due to the ambiguity of matrix clauses, one analysis of them that the learner could adopt is also one that would result in non-final finite verbs in embedded clauses. The learning model proposed here captures the influence of matrix clauses and capitalizes on their ambiguity in a principled way to arrive at an intuitive account of misanalysis in the form of parameter mis-setting.

Further, we can also note the variable production of the children as they approach adult-like performance. That is, there is a stage of development when some child productions are like (1a), while others are like (1b). A probabilistic model of learning is well suited to model such variation. As discussed in Chapter 1, we can model which grammar a speaker might use to produce an utterance on the basis of parameter values having different probabilistic weights. A learner's variable errors result from there being sufficiently heavy weights on competing parameter values.

Accordingly, a primary goal of this chapter is to illustrate a proof-of-concept model showing how parameter interaction when learning from ambiguous input can lead to parameter mis-setting. This in turn can shed light in a principled way on an account of child production errors. A second goal is to then show how the children's developmental course can be modeled in this way in a small-scale learning model. In such a model there are a number of simplifications that have been made in attempting to model the Swiss German acquisition data. Nevertheless, to the extent that the model captures the data, it receives support for the overall approach pursued here. It is to be hoped that future expansions of the model would complement and build on the insights that the approach has for explaining the acquisition puzzle.

The requirements on the model, then, are the following. The model has to first learn a grammar or set of grammars in which non-final embedded finite verbs are possible before ultimately converging on the target grammar. Since the children gradually move toward the target grammar, the model must also accommodate competing grammars during this period of variable production. Further, the model should not learn grammars that would predict non-target productions that are not attested. The probabilistic model I propose is highly successful at meeting these requirements.

In Section 2, I review the basic facts of V2 in Swiss German, which are similar to

those of standard German. Nothing about the data I discuss is particular to Swiss German, and the input corpus in Section 4 that the model learns from is consistent with German dialects in general. This raises the expectation that the modeling results I obtain for the Swiss German children can be obtained more generally when modeling children acquiring other dialects of German. This is a point I discuss in more detail in Section 7, and the acquisition data available are suggestive in pointing toward this expectation being borne out. Nevertheless, the acquisition data currently available are too sparse to be able to draw any firm conclusions. An advantage of looking at Swiss German, and one reason to focus on it for the purposes of modeling, is that Schönenberger's corpus is noteworthy for its size, containing thousands of child productions with embedded clauses. Thus in this chapter I will focus largely on the Swiss German children described in Schönenberger's work.

The structure of this chapter is as follows. I begin in Section 2 by reviewing the basic distribution of verb placement in adult Swiss German before the describing the core facts of the child productions. In Section 3, I discuss four possible analyses of the child production errors. Three of these are argued against by Schönenberger, and I will accept her arguments. The fourth is Schönenberger's own analysis. I will adopt the core insight of Schönenberger's analysis, namely that in embedded clauses children are moving the verb to a head-initial projection. Nevertheless, Schönenberger's analysis does not address the fundamental question of this chapter: why would children raise the verb to this head-initial projection when this is not allowed in the adult grammar? To answer this question, I present my model in Section 4. I first lay out my assumptions about the syntax of the adult grammar in Swiss German and then show how a prediction of the model is that children can be pushed toward a non-target grammar via the input-driven learning process itself. I present results of running the model in Section 5 along with additional child production data that can be taken to confirm these results. I compare how other models fare with respect to the acquisition puzzle in Section 6. In Section 7, I discuss how these results fit into the larger picture of German in relation to other German acquisition studies. I conclude in Section 8 with a more general discussion of the relation between the input and learning, and in particular how with an input-driven learner, input frequency can play a vital role in parameter (mis-)setting.

2. The core data of verb placement in Swiss German

In this section I introduce the core facts concerning verb placement in matrix and embedded clauses in the adult grammar and in Schönenberger's (2001, 2008) study of child utterances. As mentioned in the introduction to this chapter, I review the basic facts of V2 in Swiss German, which are similar to those of standard German. Readers familiar with the general verb placement facts in German may wish to proceed quickly through the description of the adult grammar in Section 2.1 and on to the acquisition data in Section 2.2.

2.1 Adult grammar

Here I briefly review the basic distribution of verb placement in the adult grammar of Lucernese Swiss German. As the model I present in Section 4 does not explore the fine-grained particularities of the Lucernese dialect, the data presented here are largely similar to standard German. It is of course possible that some particular properties of Lucernese play a role in the acquisition of verb placement. However, we will see that even with a non-fine-grained approach to Lucernese, the model accounts for the children's developmental course. Refining the model could lead to even more accurate modeling of this development.

The data in this section are from the St. Galler dialect of Swiss German. This is Schönenberger's native dialect and is the dialect that Schönenberger uses to provide a general sketch of Swiss German. According to Schönenberger, the St. Galler dialect is in the same dialect group as the Lucernese dialect, and the syntactic properties discussed here are representative of Swiss German in general. All examples cited throughout this chapter are from Schönenberger (2001).

Matrix clauses have characteristic Germanic V2, in which the finite verb appears in second position. A variety of phrases can occupy the initial position, such as subjects (2a), objects (2b), or adverbials (2c):

(2) a. De Rochus hät för sini Fründ die Guezli pachet.
the Rochus has for his friends these cookies baked
'Rochus has baked these cookies for his friends.'

b. die Guezli hät de Rochus för sini Fründ pachet.
these cookies has the Rochus for his friends baked

c. för sini Fründ hät de Rochus die Guezli pachet.
for his friends has the Rochus these cookies baked

(Schönenberger: 7)

A few clause types allow an initial verb. An example of this is polar questions, illustrated in (3).

(3) Chönd eer mer helfe
can you me help
'Can you help me?' (Schönenberger: 8)

However, the finite verb does not occur in any other position (e.g. third position) in matrix clauses.

In contrast, in embedded clauses the canonical position of the finite verb is clause-final. Some examples are given in (4).

(4) a. Er isch froh [dass 'Miranda mit-em Ferdinand öppedie Schach schpilt].
he is glad that the-Miranda with-the Ferdinand sometimes chess plays
'He's pleased that Miranda sometimes plays chess with Ferdinand.'

b. De Prospero isch froh [wenn-er achli in Rue läse cha].
the Prospero is glad when-he a.bit in quiet read can
'Prospero is happy when can read a bit in peace and quiet.'
(Schönenberger: 21-22)

By canonical what is meant is the absence of any of a few, highly constrained syntactic environments that allow a non-final finite embedded verb. I will shortly review these non-canonical environments below, but for the purposes of modeling child development, I will only be concerned, as far as embedded clauses are concerned, with canonical embedded clauses like those in (4). Setting the non-canonical environments aside for the moment, the three schemas in (5) characterize canonical target finite verb placement in embedded clauses.

(5) *Target finite verb placement in embedded clauses*
a. … Complementizer Subject { V_{fin} / $V_{non\text{-}fin}$ V_{fin} }
b. *… Complementizer V_{fin} Subject (X …)
c. *… Complementizer Subject V_{fin} X …

These schemas will be compared with child productions in the following section. Crucially, regardless of whether any overt material X (e.g. a non-finite verb, an object, or adverbial) co-ocurs with the subject and finite verb, the latter must be clause-final.

Turning to non-verb-final environments, in some types of embedded clauses without a phonologically overt complementizer, it is possible for the finite verb to appear in second position in the embedded clause. We see this for example under bridge verbs (e.g. *meine* 'believe'), where this verb-second position is obligatory only when the complementizer is not overt. The reader is referred to Schönenberger (2001: 12-21) for a more detailed description of embedded clauses that allow the finite verb in second position.

(6) a. Si meint [de Rochus hät de Schlüssel vegässe].
she thinks the Rochus has the key forgotten
'She believes Rochus has forgotten the key.'

b. *Si meint [de Rochus de Schlüssel vegässe hät].
she thinks the Rochus the key forgotten has

(Schönenberger: 12)

Crucially, the only overt complementizer that allows for the embedded verb to occur non-finally (albeit with a semantic effect) is *wil* 'because'. Modulo a difference in semantic interpretation, with *wil* the finite embedded verb can occur clause-finally or in the second position after the complementizer, as in [*wil* X V_{fin} … Y] (see Schönenberger 2001: 19-21).

Second, in some other embedded clauses without an overt complementizer, the finite verb can occupy initial position in the embedded clause. This is illustrated with the conditional in (7a).

(7) a. [Häsch ka Hunger meh] denn seisch mer's.
have no hunger more then say.SG me-it
'If you're no longer hungry, just tell me.'

b. [Wenn ka Hunger meh häsch] denn seisch mer's
if no hunger more have then say.SG me-it

(Schönenberger: 11)

Again, when the complementizer is pronounced, as in (7b), the verb must be clause-final. Schönenberger (2001: 11-12) details the full range of embedded clauses that allow for an initial finite verb with a non-overt complementizer.

Third, extraposition can result in the finite embedded verb being non-final. Extraposition is most common with PPs and sentential complements; according to Schönenberger extraposition with nominals is rare and is restricted to nominals that are phonologically heavy. Thus (8a) is ill-formed because the extraposed nominal is not heavy enough, whereas the extraposition of the heavy nominal in (8b) is judged well-formed.

(8) a. # Und nochler hät-er gseh es Loch. (M: 5;00)
And afterwards has-he seen the hole
'And then he saw a hole.' (Schönenberger: 243)

b. Uf Gleiss drüü chunt aa der Intercity Zug noch Bern, Lausanne, Genf.
on platform three comes particle the intercity train to Bern, Lausanne, Geneva
'The intercity train to Bern, Lausanne, Geneva arrives at platform 3.'

(Schönenberger: 213)

I note that not all constituents can be extraposed. Thus extraposition of predicative

adjectives is not possible:

(9) *T'Hex weiss, dass s Schneewittli gsi isch truurig.
the-witch knows that the Snow White been is sad
'The witch knows that Snow White was sad.' (Schönenberger: 242)

Finally, a special kind of extraposition, Verb Raising (VR) or Verb-Projection Raising (VPR), is possible in embedded clauses. This kind of extraposition is much more common than the extraposition of nominals and can be described as extraposition of the non-finite complement of certain lexically specified verbs. An example of VR is given in (10), in which the non-finite verb *fange* 'catch' embedded under the modal *wöt* 'wants' occurs clause-finally.

(10) De Ishmael weiss [dass de Ahab de wiis Waal wöt fange].
the Ishamel knows that the Ahab the white whale wants catch
'Ishmael knows that Ahab wants to catch the white whale.' (Schönenberger: 25)

In VPR, a larger constituent has extraposed rightward. In (11) this includes the embedded non-finite verb and its direct object.

(11) De Ishmael weiss [dass de Ahab wöt de wiis Waal fange
the Ishamel knows that the Ahab wants the white whale catch
'Ishmael knows that Ahab wants to catch the white whale.' (Schönenberger: 26)

The details of VR/VPR, both descriptive and theoretical are a complex topic, and I will not attempt to address them in any detail here (for an overview, see Wurmbrand 2005). For example, the optionality of these processes depends on the dialect and the embedding verb. What is important for the discussion of the acquisition of verb placement is that a number of elements are excluded from appearing in the raised cluster to the right of the finite verb. For example, subjects are excluded from appearing in the raised cluster. I discuss the significance of this in Section 3.3 in light of the fact that children produce embedded clauses in which the subject does appear to the right of the finite embedded verb.

In anticipation of the discussion of the child production errors, in which the finite embedded verb is non-final, one might wonder whether the learner has mistakenly overgeneralized one of the these four non-canonical strategies for producing non-verb-final embedded clauses. Schönenberger and I argue against the learner making such an overgeneralization; this is discussed under the alternative analyses in Section 3.1-3.3.

2.2 Child productions

Schönenberger (2001) is a detailed acquisition study of two children learning the Lucernese dialect of Swiss German. Supplementary data was gathered for Schönenberger (2008). In this section, I present a summary of the core aspects of the acquisition data, reserving discussion of analysis for Sections 3 and 4. Schönenberger's data was collected by recording the spontaneous speech of the children for a total of 150 hours over the course of a little over 4 years from ages 3;10-8;01. The study is noteworthy for collecting a rich amount of data, in particular embedded constructions, for which there is typically a small sample size of spontaneous productions in other acquisition studies. For one of the children, Moira, Schönenberger collected a corpus of approximately 5000 embedded clauses, and for Eliza (who is Moira's friend), there are approximately 600 embedded clauses.

As Schönenberger's study focuses primarily on the production of embedded clauses, there is a much smaller sample size for matrix clauses that are analyzed by Schönenberger. After extensively documenting the children's embedded clauses, Schönenberger samples matrix clauses for analysis from three periods of Moira's development (ages 3;10, 4;11, and 6;0), each sample containing several hundred clauses. I summarize Schönenberger's core findings concerning verb placement in matrix clauses before considering embedded clauses.

Moira's matrix clauses are highly consistent with the target grammar. In declaratives, the finite verb largely appears in the second position, as in the adult grammar. Schönenberger reports only a handful of examples in which the finite verb appears in third position, which is not compatible with the target grammar. With respect to other grammatical properties, Schönenberger notes that Moira's matrix clauses pattern with adult productions. For example, Schönenberger finds that word order is largely correct, with no errors after age 4;11. Further, in the first sample at age 3;10 Moira produces matrix declaratives with a range of different initial constituents, as in the adult grammar. Of 296 matrix declaratives, 168 (57%) are subject-initial, and the initial constituent in non-subject-initial clauses includes internal arguments, adverbials, and predicates. This compares similarly with findings of adult usage in Lightfoot (1997: 265), who in a corpus survey of V2 Germanic languages reports an occurrence of subject-initial matrix clauses approximately 72% of the time. Thus Schönenberger concludes that Moira's matrix clauses are largely error free.

Moira and Eliza's production of embedded clauses can be characterized by two developmental stages. In Stage 1 the children have consistent non-target verb placement in a way to be made precise below. Stage 2 is a transition from this early period to the final state. In Stage 2, the errors from Stage 1 begin to be replaced by target-like embedded clauses. This development occurs gradually throughout the course of Stage 2, with the children producing fewer and fewer errors. By the end of Schönenberger's study, the children are approaching adult-like performance.

The errors in embedded clauses in Stage 1 can be characterized as follows. Recall from (5) that in the adult grammar, the presence of an overt complementizer canonically results in a clause-final finite embedded verb. The relevant schemas are repeated below.

(5) *Finite verb placement in embedded clauses in target grammar*
a. … Complementizer Subject (… X …) V_{fin}
b. *… Complementizer V_{fin} Subject (X …)
c. *… Complementizer Subject V_{fin} X …

However, in the child productions the finite verb is frequently non-final. The schemas in (12) represent the core types of embedded clause productions in Stage 1 with respect to finality/non-finality of the finite verb. (In (12), I have abstracted away from productions in which the children have VR/VPR that is consistent with the target grammar.) Whenever some other constituent X co-occurs with the subject, the finite verb almost always occurs to the left of X, either before or after the subject (12a, b). Thus in Moira's corpus, which forms the bulk of the data, between the ages of 3;10-4;04, out of over 250 embedded clauses, the finite verb follows X only 3 times, or less than 1% (cf. Table 3.1 below). If there is no other constituent X, then the finite verb is placed either before or after the subject (12c, d).

(12) *Verb placement in embedded clauses in Stage 1 of child productions*
a. %… Complementizer Subject V_{fin} X …
b. %… Complementizer V_{fin} Subject X …
c. %… Complementizer V_{fin} Subject
d. … Complementizer Subject V_{fin}

Only the pattern in (12d) is attested in the adult grammar, and between the ages of 3;10-4;04, Moira has only one production like (12d). Further, the non-target patterns in (12b, c) are consistently produced during this period, whereas as stated above, their target counterparts are almost never attested.

In sum, verb-finality essentially only occurs if there is just a subject. Verb non-finality almost always occurs if some other constituent co-occurs with the subject. And verb non-finality can still occur if there is just a subject. The children have not yet learned the adult grammar.

One might wonder, then, whether grammatical utterances like (12d) arise via the children's adopting the adult grammar, or whether they merely look like grammatical utterances under some non-target grammar. Anticipating the analysis in Sections 3 and 4 of the errors in (12a-c), which are attributed to verb movement to a non-target head-initial position, I will follow Schönenberger in treating examples like (12d) as ambiguous. That is, it is ambiguous as to whether embedded SV clauses arise via the adult grammar with movement to a head-final position, or via a non-target grammar with movement to a head-initial position. Given that the non-target grammar accounts for all the data in the

children's early embedded clauses (cf. Sections 3 and 4), I will pursue the hypothesis that utterances like (12d), at least early in the children's development, do not arise via the target grammar.

Examples of the children's productions of the kinds of schemas in (12) are given below.

(13) a. #… Complementizer Subject V_{fin} X …
…[dass-er isch es eis] (M: 4;08)
that-he is a one
'…that he only [weighs] one (kilo)' (Schönenberger: 88)

b. #… Complementizer V_{fin} Subject X …
Prima [dass machsch du dat]. (E: 4;10)
great that make you that
'It's great that you do that.' (Schönenberger: 88)

c. #… Complementizer V_{fin} Subject
Die Ohre [wo händ eer] gfallet mer au ned. (M: 4;02)
the ears that have you.pl please me also not
'The ears which you have I don't like either.' (Schönenberger: 111)

d. … Complementizer Subject V_{fin}
Naa, isch ned luschtig [wenn ich verlüre]. (M: 4;09)
no, is not funny when I lose
'No, it's not funny when I lose.' (Schönenberger: 84)

In Stage 2 of the children's development, the children still produce the same kinds of errors that are illustrated in (12) and (13), but these are now accompanied by an increasingly frequent number of target productions of all the schemas, in particular the non-ambiguous schemas from (12a, b). Examples of target productions that correspond to (12a, b) are given below.

(14) a. Weisch du [dass das root isch]? (E: 5;01)
know you that this red is
'Do you know that this is red?' (Schönenberger: 90)

b. Deför cha si öppis [wo di ander ned cha]. (M: 5;01)
instead can she something which the other not can
'But she can do something which the other one cannot do.' (Schönenberger: 112)

In Stage 2, target production of embedded clauses gradually becomes more frequent, while non-target production becomes less frequent.

Let us now take a closer look at the developmental course graphically. Table 3.1 illustrates the developmental course of Moira's productions of embedded clauses.

Moira's corpus will be used here because its much larger size allows for a clearer illustration of the developmental trends, although comparable results hold for Eliza's productions as well. Table 3.1 is based on Tables 7.1 and 7.2 in Schönenberger (2008: 118) and shows the development of finite verb placement in embedded clauses with phonologically overt complementizers that canonically require a clause-final verb. The table represents nearly all of Moira's productions of embedded clauses introduced by these complementizers (productions in which the subject was omitted were excluded). The columns can be understood in terms of the schemas from (12) above. The column for non-target productions corresponds to the non-target schemas in (12a-c), in which the verb is not clause-final. Target productions correspond to utterances with a clause-final verb that follows the subject and some other constituent (i.e. utterances like those in (14), which are the grammatical counterparts of (12a, b)). Finally, ambiguous productions are utterances in which the finite verb is preceded only by the subject, as in (12d), or are often cases where it is possible that the children are producing target-like or relatively target-like VR/VPR or extraposition (cf. Schönenberger 2001: 66).[1]

Table 3.1 Moira's embedded verb placement productions over time

Age	**Non-target**	**Target**	**Ambiguous**	**Total**
3;10-4;04	98.44% (253)	1.16% (3)	0.38% (1)	257
4;05-4;11	78.55% (403)	2.72% (14)	18.71% (96)	513
5;00-5;05	28.30% (95)	58.52% (261)	20.17% (90)	446
5;06-8;01	6.96% (101)	70.29% (1020)	22.74% (330)	1451

[1] Schönenberger (2008: 103-106) discusses how later in development Moira appears to be applying Bernese-style VR/VPR in a relatively small proportion of productions. In Bernese Swiss German, a finite non-modal auxiliary that takes a participial complement can optionally have VR/VPR. This is not possible in Lucernese Swiss German, which Moira is acquiring, although VR/VPR is optionally possible in Lucernese when such an auxiliary takes an infinitival complement. Some of Moira's non-target, Bernese-style VR/VPR productions are classified by Schönenberger (2008: 105, Table 6) as ambiguous or target-like with respect to the error patterns in (12a-c). That is, 83 embedded clauses are taken to be ambiguous between involving the non-target verb movement pattern in (12a-c) and not involving those verb movement patterns but still having non-target Bernese VR/VPR. 70 embedded clauses are taken to be target-like in not involving the non-target patterns in (12a-c), even though they unambiguously contain the non-target Bernese VR/VPR pattern. In addition to these 153 tokens, which occur between ages 5;00-8;01, an additional 22 tokens between ages 3;10-4;11 are classified as ambiguous like the 83 tokens mentioned above. Even though Schönenberger does not consider all these kinds of examples to be errors of the types in (12a-c), it does not seem unreasonable to me to classify them as such. However, it is unclear to what extent these 175 tokens are included in Table 3.1 under the 'Target' and 'Ambiguous' columns. Should they all be added to the 'Non-target' column, then Moira's error rates would increase slightly.

Examining Table 3.1, we can see that the in the first months of the study there are virtually no target-like counterparts to (12a-c). This is Stage 1. A few more scattered instances of target counterparts to (12a, b) begin to occur at ages 4;05-4;07. These early target productions first occur with the more commonly used complmentizers *dass* 'that', *wenn* 'when/if', and *öb* 'whether/if' (cf. the detailed graphs in Chapter 2 of Schönenberger 2001). This is the beginning of Stage 2. At first these target productions are relatively infrequent, but there is an upward trend in their use, especially after age 4;11. By the end of the study they have largely displaced their non-target counterparts.

Table 3.1 clearly illustrates the learning trajectory to be replicated by the model: a period of high error rates from the recorded beginning of productions, followed by gradually diminishing error rates.

3. Some possible analyses

Schönenberger (2001) suggests three alternative analyses of the child production errors before proposing her own. Schönenberger identifies empirical challenges for these alternatives. I review these alternatives in Sections 3.1-3.3. These analyses all have challenges in accounting for the full range of data attested in the corpus of child productions. This is not to say that the arguments against these analyses are decisive; rather I wish to flag the difficulties that they face in accounting for the data in a principled way. I then give the basics of Schönenberger's own analysis in Section 3.4. The core idea is that child production errors involve non-target verb-raising to a head-initial projection. However, this analysis suffers in that it provides no explanation for why children would adopt this non-target grammar. Nevertheless, I adopt the core of this hypothesis and show in Section 4 how the learning process itself of the model I propose can initially push the learner toward this non-target grammar.

3.1 Alternative #1: Overgeneralizing V2 in embedded clauses

An intuitive possibility to pursue is that the unexpected leftward position of the finite verb in child productions results from children failing to distinguish between matrix and embedded clauses. According to this hypothesis, children have correctly learned V2 (as well as verb-initiality) in matrix clauses, but they have not yet learned to treat embedded clauses as any different from matrix clauses. Consequently they incorrectly overgeneralize V2 found in matrix clauses to embedded clauses. Nevertheless, this hypothesis suffers from two shortcomings.

First, regardless of how one defines the domain for V2 in child embedded clauses, the V2 hypothesis cannot account for all child productions. I will consider two variants of an embedded V2 analysis. Under the first variant, the domain for V2 in embedded clauses is defined as following the complementizer. This analysis of embedded V2 arises under the proposal of CP-recursion, which has been proposed for some West Germanic

languages (see, for example, de Haan and Weerman 1986). In CP-recursion, the complementizer occupies a higher CP, which embeds a lower CP. If the children are generalizing embedded V2 with a CP-recursion structure, then the finite verb will appear in the lower C-head, and some other constituent will appear in SpecCP of the lower CP. This means that the finite verb should never appear as the third (or greater) constituent after complementizer. This holds in productions such as (13a-c) repeated below; however, this is not the case in (15), where the finite verb appears in fourth position after the complementizer.

(13) a. #… Complementizer Subject V_{fin} X …
…[dass-er isch es eis] (M: 4;08)
that-he is a one
'…that he only [weighs] one (kilo)' (Schönenberger: 88)

b. #… Complementizer V_{fin} Subject X …
Prima [dass machsch du dat]. (E: 4;10)
great that make you that
'It's great that you do that.' (Schönenberger: 88)

c. #… Complementizer V_{fin} Subject
Die Ohre [wo händ eer] gfallet mer au ned. (M: 4;02)
the ears that have you.pl please me also not
'The ears which you have I don't like either.' (Schönenberger: 111)

(15) # [wenn du no einisch wärsch do drin] (M: 4;11)
if you still once would-be there in
'if you were still once in there' (Schönenberger: 307)

Recall from Section 2.2 that the finite verb occurs in third position in Moira's matrix clauses only a handful of times, and only up until age 4;11. In contrast, there are at least 22 examples of embedded clauses in Moira's corpus of the type in (15), most of which are attested after age 4;11 (Schönenberger 2001: 146-147). This result is unexpected under the V2 analysis of embedded clauses. Under this version of the embedded V2 analysis, we expect the verb in (15), for example, to precede *einisch* 'once', contrary to its actual position in (15), which follows this constituent.

Under the second variant, if the domain for V2 in embedded clauses is defined as including the complementizer, then the finite verb should never appear as the second or third constituent after the complementizer. This would account for examples such as (13b, c) but would not account for the hundreds of productions like those in (13a) or those in (15), where the finite verb is preceded by at least the subject.

Second, Schönenberger presents some additional evidence showing that Moira's embedded clauses pattern differently from her matrix clauses. As was discussed in Section 2.2, Moira's matrix clauses are adult-like in allowing for a range of non-subject

constituents to appear in initial position before the finite verb 43% of the time. However in Moira's embedded clauses, other than the complementizer, non-subjects only rarely precede the finite verb. Schönenberger compares the production of matrix clauses to that of embedded clauses introduced by the complementizers *wenn* 'when, if', *dass* 'that', and *öb* 'whether, if'. Of 179 productions of embedded clauses introduced by these complementizers in which there is an overt constituent between the complementizer and the finite verb, that constituent is a non-subject only 4 times (2.2%), and none of these constituents is a nominal argument. 2 of these examples have the adverbial *etz* 'now' preceding the finite verb, and the other 2 examples involve discourse particles, such as *de* 'then' (Schönenberger 2001: 254). This suggests that Moira distinguishes between matrix and embedded clauses. In order to maintain the V2 analysis of embedded clauses, one would need to explain how children have failed to distinguish the two types of clauses with respect to V2 but have learned to distinguish them with respect to their initial constituent. I will not pursue this possibility and leave it as an open challenge to the V2 analysis.

There is a third variant of the V2 analysis, which is based on the possibility of certain types of embedded clauses allowing V2 (or verb-initiality) in the adult grammar. Recall from Section 2.1 that in certain embedded clauses V2 arises without a phonologically overt complementizer. Indeed, when the children produce these types of clauses without an overt complementizer, the finite verb occurs non-finally as per the schemas in (12) from Section 2.2. An example of this given below (cf. the adult (6a)).

(16) Etz han-ich gmeint [du chöntisch ned Rinde ässe weg de
now have-I thought you could not rind eat because of.the
Schpange]. (M: 4;08)
braces
'I thought you couldn't eat any rind because you're wearing braces.'
(Schönenberger: 76)

According to this variant of the V2 analysis, one could say that the children fail to distinguish between embedded clauses with a complementizer and those without that allow non-verb-finality. Children would have correctly learned non-verb-finality in non-canonical complementizer-less embedded clauses, but they have not yet learned to treat canonical embedded clauses as any different from these non-canonical embedded clauses without overt complementizers. Consequently they incorrectly overgeneralize V2 found in one kind of embedded clause to all embedded clauses In this new analysis, the overt complementizer would not be considered part of the domain for V2.

Again, this variant is subject to similar challenges as the original V2 analysis of embedded clauses. This new analysis accounts for the types of productions in (13a-c), but fails to account for examples of the sort in (15) for the reason discussed above.

There is also some slight evidence showing that children distinguish between the two types of embedded clauses based on whether the initial, pre-verbal constituent is a

non-subject. Recall that in the sample of Moira's embedded clauses with an overt complementizer that are canonically verb-final in the adult grammar, a non-subject constituent occurs between the complementizer and the finite verb only 2.2% of the time, and none of these constituents is a nominal argument. In 191 productions of embedded clauses introduced by a bridge verb without a complementizer in Moira's corpus, 15 productions (7.9%) occur with a non-subject constituent in initial position between the bridge verb and the finite embedded verb. Further, these 15 productions are roughly split in half between having nominal and non-nominal initial constituents, exhibiting a greater range of initial constituents than in the sample mentioned above of Moira's embedded clauses with a complementizer. The sample sizes here are smaller, and so the evidence is less compelling, but taken as a whole it suggests that the children treat embedded clauses with overt complementizers as different from other types of clauses. If one wanted to pursue this third variant, there is now a learning challenge similar to the one raised by the original V2 analysis: how to account for children failing to distinguish across embedded clause types when it comes to verb placement, but distinguishing across them with respect to initial constituents.

I assume, then, that verb non-finality in embedded clauses that canonically verb-final in the adult grammar is not the result of the children having incorrectly overgeneralized V2 (or V1) patterns from other types of clauses. Two natural alternatives present themselves to account for child non-target verb placement. These are the additional two syntactic phenomena from Section 2.1 that can result in a non-final embedded clause, namely extraposition and VR/VPR. In the following two subsections I will reject alternative analyses that make use of these phenomena to account for the child production errors.

3.2 Alternative #2: Extraposition in embedded clauses

In Section 2.1 we saw that another strategy available in the adult grammar for non-verb-final embedded clauses is extraposition. Here I consider an alternative analysis of child production errors that relies on extraposition (of the non-VR/VPR variety). According to this alternative, children have overgeneralized the use of extraposition, resulting in non-target embedded clauses. However, this analysis raises several challenging questions.

First, Schönenberger reports that in Moira's matrix clauses there are only a handful of productions involving non-target-like extraposition of phonologically light nominals. Further, recall from Section 2.1 that extraposition of predicative adjectives is not possible in the adult grammar. The sample of Moira's matrix clauses is target-like in not containing any extraposed predicative adjectives. Although the number of productions with matrix predicative adjectives that are not extraposed is not reported, Schönenberger (2001: 196) provides at least one clear example of target-like non-extraposition with a predicative adjective in a matrix clause. Although not conclusive, the evidence discussed above is suggestive that Moira has learned target-like extraposition in matrix clauses. I

will assume that this is largely correct. In contrast, in 70 examples in Moira's corpus of embedded clauses introduced by *wenn* 'when, if', the finite verb is followed by a predicative adjective, as illustrated below.

(17) # [Wenn si isch heiss] denn isch si besser. (M: 3;11)
when/if she is hot then is she better
'When/if it [the soup] is hot, then it's better.' (Schönenberger: 242)

We are now faced with a familiar learning question: how could the children putatively learn target-like extraposition in matrix clauses but fail to do so in embedded clauses?

Second, even if further investigation found some more extraposition errors in matrix clauses, the current sample of matrix clauses shows that these errors are not systematic in matrix clauses, in contrast to the systematic non-target productions in embedded clauses. If one were to pursue an extraposition analysis, one faces the question as to why extraposition rates are so high in embedded clauses. This discussion does not argue definitively against an extraposition analysis, but I leave these questions as open challenges to this kind of analysis.

3.3 Alternative #3: Overgeneralizing VR/VPR

As a final alternative, one could pursue an analysis of production errors by claiming they involve VR/VPR extraposition, which is yet another strategy available in the adult grammar for non-verb-final embedded clauses, as discussed in Section 2.1. Examples of these strategies are repeated below from Section 2.1.

(10) *VR*
De Ishmael weiss [dass de Ahab de wiis Waal wöt fange].
the Ishamel knows that the Ahab the white whale wants catch
'Ishmael knows that Ahab wants to catch the white whale.' (Schönenberger: 25)

(11) *VPR*
De Ishmael weiss [dass de Ahab wöt de wiis Waal fange
the Ishamel knows that the Ahab wants the white whale catch
'Ishmael knows that Ahab wants to catch the white whale.' (Schönenberger: 26)

According to this analysis, children would have correctly learned to apply VR/VPR to the non-finite complements of VR/VPR verbs, but they have failed to distinguish between VR/VPR verb and non-VR/VPR verbs and overgeneralize by applying VR/VPR consistently to all verbs in embedded clauses regardless of their complement (whether the complement be verbal at all, infinitival, or participial).[2]

[2] It is still possible that Moira has done *some* over-generalization of VR/VPR, especially after unambiguously target-like productions of VR/VPR have been attested. This possibility was pointed out in note 1 for a small subset of the data. The point in this section is that it seems unlikely that

One difficulty with this analysis is that there is not clearly an independent way of verifying whether children have learned VR/VPR. Unlike, for example, the extraposition in Section 3.2 for which we can examine target productions in matrix clauses, VR/VPR is strictly an embedded clause phenomenon. Minimally, then, if children have learned VR/VPR but simply overgeneralize which verbs it applies to, we would expect their embedded clauses to conform to other properties of VR/VPR. As was mentioned in Section 2.1, certain constituents cannot appear in the raised cluster that follows the finite verb in VR/VPR environments. Notably this includes the embedded subject. Thus if children have correctly learned VR/VPR, then we do not expect embedded clauses where the subject follows the finite verb. Nevertheless, this occurs hundreds of times in Moira's corpus; cf. (13b):

(13) b. #… Complementizer V_{fin} Subject X …
Prima [dass machsch du dat]. (E: 4;10)
great that make you that
'It's great that you do that.' (Schönenberger: 88)

In its simplest form, then, the VR/VPR analysis does not account for all production errors.

One could attempt to modify the VR/VPR analysis by claiming that children have overgeneralized the range of constituents that can appear in the raised cluster. But this only raises further questions. For instance, one could ask why children do not overgeneralize VR/VPR to matrix clauses. As Schönenberger notes, if children incorrectly apply VR/VPR in embedded clauses with the result of having a subject in the raised cluster, then we might expect similar errors in matrix clauses. Such a non-target matrix clause would look like (18), but such errors do not appear to be attested in Moira's sample of matrix clauses (cf. Schönenberger 2001: 241).

(18) XP V_{fin} $V_{non\text{-}fin}$ Subject …

To maintain the VR/VPR analysis, one is left with the question of how children learn some properties of VR/VPR (e.g. extraposition past the finite verb), but not others (e.g. exclusion of the subject in the raised cluster; only certain lexically speicified verbs allow this raising; etc.). I leave this as a further open challenge and turn to Schönenberger's own analysis, which I develop further in Section 4.

---

overgeneralization of VR/VPR could account for all the verb-placement errors, especially those that do not even involve an auxiliary.

3.4 Schönenberger's analysis: Verb movement in embedded clauses

In Schönenberger's (2001) analysis, the children's non-target embedded clauses involve raising the verb to a higher position. A strength of this analysis is that it can account for all the production errors we have seen. In what follows I present a simplified version of this analysis. This simplification does not detract from the scope of the empirical coverage of Schönenberger's analysis; nor does it fundamentally change the nature of the learning questions it raises.

Schönenberger assumes that in matrix clauses in the target grammar, the finite verb raises to a C-head. Following Rizzi (1997), Schönenberger assumes there are a number of C-heads in the clausal periphery, but it is sufficient to say that in the target grammar, the finite verb raises to a high position in the C-domain. In accordance with the Head Movement Constraint (Travis 1984), the verb must move through all intermediate head positions. In embedded clauses, the overt complementizer occupies this high position and blocks the finite verb from moving there.

In the child grammar Schönenberger first assumes that the children have learned to raise the verb to a C-head in matrix clauses. This accounts for their largely target-like production of matrix clauses. Second, Schönenberger assumes that the children have learned that the complementizer is a head occupying some high position in the C-domain. Accordingly, verb movement is blocked to that position. Instead, children raise the finite verb to some head position outside of the VP but below the complementizer. Let us call this head H.[3] As the finite verb in non-target productions regularly precedes non-subject constituents, H must be head-initial. This is shown schematically below.

(19) … dass (Subject) [$_{HP}$ V+H [(Subject) X ~~V~~]]

This analysis is able to account for all the child productions of embedded clauses that we saw in (12).

(12) *Verb placement in embedded clauses in Stage 1 of child productions*
a. #… Complementizer Subject V_{fin} X …
b. #… Complementizer V_{fin} Subject X …
c. #… Complementizer V_{fin} Subject
d. … Complementizer Subject V_{fin}

First, as the verb is raising to some higher position, it is possible for constituents to follow it (12b). Second, given the well-known relative freedom of word-order in the middle-field in German (cf. Haider 2010), so long as we assume that the subject can

[3] Schönenberger, in fact, assumes that there are multiple heads between the complementizer and the VP that the verb can target when it moves. A primary motivation for this concerns more detailed data involving stressless subject pronouns, which I discuss in Section 5.3. If one assumes an analysis of head-adjunction for these pronouns, as I do, then the motivation for raising the verb to these additional positions disappears. See Schönenberger (2001: 286-304) for details.

appear in multiple positions, it could precede or follow the finite verb (12b-d) (cf. Diesing 1992).[4,5] Finally, adjunction to HP allows for the possibility of some additional constituent to precede the finite verb as we saw in (15), repeated below.

(15) # [wenn du no einisch wärsch do drin] (M: 4;11)
if you still once would-be there in
'if you were still once in there' (Schönenberger: 307)

Data such as (15) are less common in child productions and appear to be a later development (cf. Section 5.3). We can interpret this as adjunction to HP being an option that children do not make much use of, especially in earlier developmental stages, and to the extent that they might make use of it, it is largely confined to adjunction of subjects. I have no explanation for why this might be the case, but this claim appears to be borne out in the more detailed analysis of the acquisition data in Section 5.3. There we will see that at least initially, the children have a tendency to not have material in SpecTP or adjoining to TP (which I identify as HP). Accordingly, we do not expect massive adjunction to HP to obviate errors of the type in (12a, b).

Schönenberger's analysis is appealing for several reasons. First, it provides a clear account of the core patterns of embedded clause productions. Second, unlike the extraposition and VR/VPR analyses considered above, this analysis of embedded clauses does not predict any additional errors in the children's matrix clauses. If the children are raising the verb to a C-head in matrix clauses, then they are already raising it via movement to the head H. Thus movement to H in embedded clauses does not result in anything that is not already attested in matrix clauses. Third, for the reason just discussed, the analysis draws a clear connection between learning in matrix and embedded clauses. It allows for the possibility of the child transferring what has been learned in matrix clauses to embedded clauses. In Section 4 below I will make precise why this kind of transference is predicted in the learning model I propose via parameter interaction.

Nevertheless, Schönenberger's analysis raises questions of its own about the learning process. Two key questions arise. First, how is it that children adopt a grammar in which the functional projection HP that the embedded verb raises to is head-initial? As Haider (2010) discusses, there is no clear evidence of head-directionality in the middle-field in German. This ambiguity will be a crucial component in my account of the errors, which I discuss in Section 4. In Schönenberger's analysis the ambiguity arises because in the adult grammar, the finite verb always raises past the middle-field in matrix clauses; thus H could be head-initial (20a) or head-final (20b):

[4] I will use the term middle-field as a descriptive term for whatever syntactic structure might appear in the adult grammar between the base position of any verbs and the position of either (a) the complementizer, or (b) the finite verb in matrix clauses or V1/V2 embedded clauses, such as those without an overt complementizer.

[5] The children also make use of the non-target option of having a stressless pronoun occur after a non-final finite verb in embedded clauses, as in (13b). See note 28 for some speculative remarks about this.

(20) a. [$_{CP}$ XP [V+H+C [$_{HP}$ ~~V+H~~ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]]]]]
b. [$_{CP}$ XP [V+H+C [$_{HP}$ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]] ~~V+H~~]]]

And in embedded clauses in the adult grammar, the verb could appear in-situ in the VP or could raise to the middle-field. Under this view, case-licensing of the subject, whether in-situ or not, could proceed via agreement with T and would not be dependent on the subject moving to TP (cf. Miyagawa 2001). With an in-situ subject, any projection in the middle-field could be head-initial (21a) or head-final (21b). If the subject moves to the middle-field, the head H would have to be head-final (22a) so as to maintain verb-finality (cf. (22b)).

(21) a. Subj V [$_{CP}$ dass [$_{HP}$ H [$_{VP}$ Subj [$_{VP}$ Obj V]]]]
b. Subj V [$_{CP}$ dass [$_{HP}$ [$_{VP}$ Subj [$_{VP}$ Obj V]] H]]

(22) a. Subj V [$_{CP}$ dass [$_{HP}$ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]] V+H]]
b. *Subj V [$_{CP}$ dass [$_{HP}$ V+H [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]]]]

Thus it is not clear in Schönenberger's analysis how the child would first learn a grammar that consistently has a head-initial projection in the middle-field.

Second, Schönenberger does not address how the child arrives at the target state. In other words, if a grammar involving non-target verb movement as in (19) is viable for the child, how does the learner move from the non-target grammar and converge on the target grammar? [6]

In the analysis I present in Section 4.1, I adopt Schönenberger's core claim that the production errors result from non-target verb raising to some head-initial projection in the middle-field. My analysis is somewhat different, and accordingly the learning questions it raises differ somewhat from those just presented, as will be discussed.

4. A learning model for the acquisition puzzle

In this section I present the basics of an implementation of the learning model in Chapter 2 that can address the acquisition puzzle. I begin by laying out my assumptions regarding of the adult and child grammars in Section 4.1, before presenting an overview of the model in Section 4.2. The insight of the learning model follows in Section 4.3, and predictions for the model's results are given in Section 4.4.

4.1 Analysis of the adult and child grammars

In this section I present an analysis of the target adult grammar and then the non-target child grammar. Detailed discussion of why the learning model predicts the children to

[6] In fact, Schönenberger (2001) does not appear to have any clear stance on what the target grammar looks like and is apparently agnostic between both (21) and (22a) (cf. Schönenberger 2001: 289).

first learn a non-target grammar can be found in Section 4.3 after the learning model is introduced in Section 4.2.

For the adult grammar the analysis I adopt is based on what Zwart (1997) and Haider (2010) call the traditional generative analysis of V2 Germanic languages. This originates in the work of Bach (1962), Koster (1975), and Den Besten (1977/1989) among others. A clear exposition of this kind of analysis in the Principles and Parameters framework (Chomsky 1981) can be found in Platzack (1986). A primary reason for pursuing this kind of analysis is that it allows for a clear illustration of how the learning mechanism introduced in Chapter 1 (i.e. parameter interaction) leads in a principled way to the following situation of transference across clause types: learning from matrix clauses affects what is learned about embedded clauses. In Section 4.3 I show how this plays an important role in predicting child errors.[7]

According to the traditional analysis, the V2 property of matrix clauses in Swiss German results from the finite verb moving to C, which is head-initial. Whenever the verb appears in second position, this verb-movement co-occurs with movement of some other constituent to SpecCP.

Verb movement can be implemented in the grammar by setting 2 head-movement parameters. One is a V-to-T parameter [±VT], a positive setting of which requires the finite verb to move to T if no free morpheme blocks it from doing so. The second is a T-to-C parameter [±TC], a positive setting of which requires T (including the V+T complex) to move to C if no overt complementizer blocks it. I present these parameters formally here:

(23) *T-to-C movement parameter*
  a. [+TC]: T moves to C if no free morpheme in C blocks it
  b. [–TC]: T does not move to C

(24) *V-to-T movement parameter*
  a. [+VT]: finite V moves to T if no free morpheme in T blocks it
  b. [–VT]: finite V does not move to T

In other words, the V2 property results from 2 independent steps of movement that are obligatory whenever possible (i.e. not morphologically blocked). There is obligatory V-to-T movement in all clauses, and there is obligatory T-to-C movement whenever there is no complementizer. Given this hypothesis space and the Head Movement Constraint, it follows that if only one of these parameters is positively set, then the verb cannot be in C. It also follows that whenever there is evidence for verb movement at all, then that input is unambiguous evidence for a [+V-to-T].

Further, the position of the verb can be determined by head-directionality

[7] There are other analyses of German, perhaps most prominently Zwart (1997). As a topic for future research I leave the question of how well a model based on Zwart's analysis and syntactic parameters would fare with modeling the developmental course of the Swiss German children.

parameters. In addition to C being head-initial we can ask whether T and V are head-initial or head-final, [T-init/fin] and [V-init/fin] respectively. Thus I do not assume that all projections are head-initial, *contra* Kayne (1994). The parameters are presented below.

(25) *CP-headedness parameter*
a. C-in(itial): C linearly precedes its sister complement
b. C-fin(al): C linearly follows its sister complement

(26) *TP-headedness parameter*
a. T-in(itial): T linearly precedes its sister complement
b. T-fin(al): T linearly follows its sister complement

(27) *VP-headedness parameter*
a. V-in(itial): V linearly precedes its sister complement
b. V-fin(al): V linearly follows its sister complement

A key observation to make is that verb movement to C results in ambiguity for the headedness of TP in all matrix clauses. I alluded to this ambiguity in the previous section in the context of Haider's (2010) discussion of the middle-field in German. To see this ambiguity, consider the following. In the hypothesis space above, verb movement to C requires a [+VT, +TC] grammar, but as the V+T complex always vacates TP, the headedness of TP is entirely ambiguous in matrix clauses – both structural analyses are compatible with matrix clauses. This is illustrated schematically in (28), which replaces H from (20) with T.[8]

(28) a. [$_{CP}$ XP [V+T+C [$_{TP}$ ~~V+T~~ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]]]]]
b. [$_{CP}$ XP [V+T+C [$_{TP}$ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]] ~~V+T~~]]]

In fact, the ambiguity for TP-headedness is not dependent on the particulars of the hypothesis space I have proposed. So long as the learner can consider the hypothesis that the finite verb is in C in matrix clauses, regardless of the exact formulation of the verb movement parameter(s) is, the basic ambiguity in (28) persists. Regardless of what one's analysis of German is (e.g. Zwart 1997 does not assume that the finite verb is always in C), it seems to me implausible to exclude from the learner's consideration the hypothesis that the verb is in C given that such a possibility is widely acknowledged as being allowed by Universal Grammar. Thus I take very seriously the claim that there is a high degree of ambiguity for TP-headedness in (Swiss) German, and an important consideration for the learning model, discussed in Section 4.3, is this ambiguity that we

[8] I note that this kind of analysis for V2 needs to make the assumption, which is often implicit in the literature, that CP is privileged in not allowing some other constituent YP to also adjoin to CP in (28). Such a configuration would incorrectly allow the finite verb to appear in third position, and so the possibility of some YP and XP co-occurring in CP is ruled out by assumption.

see in matrix clauses with respect to TP-headedness.

Given this ambiguity, what can we say about the target grammar with respect to TP-headedness in Swiss German? The issue of TP-headedness has figured prominently in the debate over the phrase structure of German since at least Travis (1984). However it is not clear that the literature provides unequivocal empirical evidence for either T-initial or T-final in German. A review of the literature that would do any justice to this debate would take us too far afield, and I refer the reader to the following sampling of works that provide contrasting viewpoints on empirical and theoretical aspects of the debate but that do not clearly resolve the basic analytical question of TP-headedness: Schwartz and Vikner (1996), Zwart (1997), Sells (2001), Meinunger (2007), and Haider (2010). That the debate does not appear to be resolved can be taken as an indication of the following two points. First, as mentioned above, that the evidence brought to bear on TP-headedness so far has been inconclusive. Second, that a new approach can be fruitful in shedding light on the issue. My proposal is that incorporating formal learning and acquisition modeling perspectives into the discussion might be able to do just that. To be clear, I will not attempt to resolve the debate here. The methodology I propose is a potential way of doing so. What I will undertake here is a first step in this approach. It will allow us to get a model off the ground and running, and it will bring some empirical results that bear on the issue. This first step is to show that the learning model provides evidence in support of a target setting of T-final given the 5-parameter hypothesis space introduced above. However, such a result is not by itself conclusive evidence against a target setting of T-initial. Such evidence would come from running the model with alternative hypothesis spaces, and the results of these simulations could be used to argue against a target T-initial setting. This kind of detailed comparison across hypothesis spaces remains an area for future research. What I will present here is simply one corner of it, but one that has the pleasing outcome of providing results that model the developmental trajectory of the Swiss German children. Let us now consider the methodology of this proposal in more detail.

A learning model, coupled with a particular set of parameters, can provide evidence in support of a given syntactic analysis of TP-headedness. The idea is to use a hypothesis space to come up with an analysis and then to test that analysis with a simulation of the learning model. There are thus two components at play: developing an analysis and evaluating that analysis with a model. As for the analysis, in a given hypothesis space, we can make a proposal about what the target grammar is based on whether that grammar accounts for the entire corpus of input. I will provide an illustration of this below, in which I show that in the hypothesis space with the 5 parameters that I introduced above, the target grammar is T-final. Different formulations of the parameter space could lead to different proposals about what the target grammar is.[9, 10] Second, evaluating the analysis

[9] Zwart (1997), for example, assumes a rather different framework for verb movement parameters and concludes that T is head-initial. As mentioned in note 7, though, it is unclear whether a learning model

proceeds as follows. Support for the analysis, and by extension the hypothesis space, lies in the extent to which a learning model can converge on the putative target grammar while still modeling children's developmental trajectory, including attested errors. Modeling results that do not converge on the putative target grammar (i.e. they converge on some other grammar, which does not account for all the input) can then be taken as evidence that the hypothesis space should be adjusted, which can in turn lead to a different analysis. (It is even possible that within a given hypothesis space, there are two analyses that account for the entire corpus of input. In such a case we can use results from the learning model to evaluate what the target grammar is. If the model consistently learns a single grammar, we can say that that grammar is the target, and if the model sometimes learns one grammar and sometimes another, that leads to the prediction that there is grammatical variation across learners; cf. note 10, and see also Chapter 4 for a more detailed illustration of this.) I will also propose that the Swiss German children's errors reflect a parameter mis-setting of T-initial. We will see that the learning model I have proposed can converge on the target T-final grammar while also being able to mis-

---

operating with Zwart's hypothesis space would be able to model the child errors in Swiss German. The extent to which it is unsuccessful in doing so can then be taken as evidence against such a hypothesis space.

[10] Another hypothesis space we might consider is one that is minimally different from that in the text. This alternative hypothesis space augments the 5 parameters in the text with a sixth parameter of V-to-C movement. This parameter provides what can be thought of as a rather intuitive account of V2 by means of a single parameter. If positively set, this parameter would require verb movement to C in case there were no complementizer. Further, verb movement would proceed cyclically through T, but verb movement to T would not necessarily be obligatory in every clause. For example, the learner might have a [+V-to-C, –V-to-T] grammar, which can account for all the Swiss German input. Under such a grammar, in matrix clauses, the verb would move to C (through T), but embedded clauses with a complementizer would block verb movement to C. As such a grammar is [–V-to-T], the verb in these embedded clauses would remain in-situ in, for example, a head-final VP, thus resulting in clause-finality of the verb in embedded clauses. We can call this proposal a V-to-C hypothesis space.

What might this hypothesis space tell us about TP-headedness? A [+V-to-C, –V-to-T] grammar can account for all the Swiss German input. But as the verb never remains in T (either moving on to C, or remaining in the VP), we might wonder whether this hypothesis space tells us anything at all about TP-headedness. That is, in this hypothesis space both a T-initial and a T-final grammar can account for all the input. After discussing how the model learns via parameter interaction in more detail in Section 4.3, I return to this question in note 21. In brief, even in a V-to-C hypothesis space we expect the model to learn a T-final parameter setting.

I note that the V-to-C hypothesis space is more complex than the one proposed in the text, but that it appears to cover the same empirical ground. The V-to-C analysis can certainly account for the adult grammar, but it is not necessary to have a V-to-C movement parameter in order to account for V2. As I have shown, V2 can be accounted for with separate V-to-T and T-to-C movement parameters. And as V-to-T and T-to-C movement parameters find independent motivation in, for example, Pollock's (1989) account of verb movement in French and English, I will assume that they are part of the learner's hypothesis space. Further, I illustrate later in this section how separate V-to-T and T-to-C parameters can account for the contrasting verb placement patterns in German matrix and embedded clauses. Thus although a V-to-C parameter provides a straightforward account of V2, the cross-linguistic motivation for it is less clear.

A perhaps more pertinent question is whether a V-to-C hypothesis space would be able to model learner errors. As I discuss in note 21, the V-to-C hypothesis space has a smaller likelihood of modeling the development of the Swiss German children, but more simulations of the model are necessary to evaluate this question. Should a V-to-C hypothesis ultimately prove unsuccessful in modeling errors, then that result can be taken as evidence that a V-to-C parameter is not part of any learner's hypothesis space.

set T-initial during the course of learning. The modeling results can thus be taken as evidence in support of the syntactic analysis of Swiss German that I pursue.

In light of the 5 parameters above, let us consider how they lead to the analysis of the adult grammar being T-final. An important assumption concerning these head-movement and head-directionality parameters is that they apply uniformly in all types of clauses. Thus, what can be concluded about one type of clause can be informative in analyzing another type of clause. It is in this sense that (mis)learning something about matrix clauses can affect what is learned about embedded clauses in the child grammar. As for the adult grammar, examining embedded clauses while assuming a [+VT, +TC] grammar allows us to conclude what the headedness of TP is for all clauses. Recall that verb movement to C in matrix clauses is possible because there is no overt complementizer that blocks it. In embedded clauses without a complementizer, the verb raises to C (cf. the examples in (6) and (7), in which the embedded verb is non-clause-final). This movement is blocked in embedded clauses with an overt complementizer. Nevertheless, in a [+VT, +TC] grammar, although T-to-C movement may be blocked, V-to-T movement is not. The analysis leads us to conclude that there is string-vacuous movement of the finite verb to T and that TP is head-final. This accounts for the verb-final property of canonical embedded clauses; this is illustrated schematically in (29), which modifies (22).

(29) a. Subj V [$_{CP}$ dass [$_{TP}$ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]] V+T]]
b. *Subj V [$_{CP}$ dass [$_{TP}$ V+T [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]]]]

If T were head-initial in embedded clauses, then V-to-T movement would result in non-target clauses there were not verb-final (29b). Thus, in the adult grammar there is obligatory V-to-T movement, and TP is head-final.

We can see that T must be head-final most clearly when certain kinds of matrix clauses co-occur with embedded clauses in the same token of input. Consider the string *[XVSO [Comp. SOV]]*. As the verb precedes the subject in the matrix clause, it must have raised outside of the VP to some head-initial projection. Minimally this input must be [+V-to-T], but does the verb move to C as well? Verb movement to C would tell us that CP is head-initial, but if the verb raised only as high as T in the matrix clause, then TP would have to be head-initial. Assuming uniformity across clauses, and given [+V-to-T], the verb must have raised in the embedded clause as well. But the embedded verb follows the object in its clause, and so it must have raised to a head-final projection. (If the embedded clause were just SV, then it would not be clear that the embedded verb raised to a head-final position.) There is only one set of parameter values in (23)-(27) that is consistent with this distribution. The finite verb in the matrix clause raises to a head-initial CP, and as this movement is blocked in the embedded clause, the embedded finite verb raises to a head-final TP. Thus the co-occurrence of certain embedded clauses with certain matrix clauses in the same input token provides unambiguous evidence for a

[+T-to-C, T-fin] grammar.

Nevertheless, as we shall see below, the vast majority of all matrix clauses do not co-occur with any embedded clause at all. These matrix clauses are thus ambiguous for TP-headedness, and whatever the model learns from this ambiguity can affect how a learner produces embedded clauses.

The discussion above, leads me to reject Lightfoot's (1989, 1991) Degree-0 hypothesis concerning language acquisition. According to this hypothesis, the child sets all parameters solely on the basis of evidence from the matrix clause and the edge of an embedded clause. In Lightfoot (1991) it is proposed that the learner has access all the way to Infl (or the T-head in the analysis here) of the embedded clause, but not the VP. What does the embedded TP tell the child learning Swiss German with the hypothesis space in (23)-(27)? If the learner only had access to the embedded TP, but not the embedded VP, why would a learner who had adopted a T-initial parameter setting ever reject such a grammar? It is only by considering the surface position of material that is structurally below TP (e.g. a VP-internal object in (29)), that a learner with a verb raising grammar can move toward a T-final grammar from a T-initial one.

Departing from Lightfoot's hypothesis is a sensible move for the probabilistic learner I have proposed. Lightfoot's motivation is to simplify the learning task for the child: the child would need to pay attention to only a simplified subset of the input. For a learner that is waiting to hear unambiguous triggers for parameter settings, the Degree-0 hypothesis is very helpful in reducing the search space for such triggers (assuming, of course, that all such triggers can be found in the Degree-0 input). Input without any triggers, though, whether in the Degree-0 input or not, is simply uninformative to the learner. In contrast, a probabilistic learner can learn incrementally from any kind of input no matter how ambiguous. Any token of input can be informative about the grammar of best fit for a particular language. As we saw in Chapter 2, a natural step to take, then, is to suppose that the model is trying to learn the grammar that best reflects *all* the input. Seen in this light, rather than making the learning task more of a challenge, the more data considered by the learner, the easier it can be for that learner to acquire the adult grammar.

So far in this section, I have discussed the following components of the adult grammar: (a) the V-to-T and T-to-C verb movement parameters are positively set; (b) CP is head-initial; and (c) TP is head-final. Before turning to my account of the children's grammar, I take up two more points of the adult grammar. The first concerns constituents in the middle-field, and the second is the parameter setting for VP-headedness.

First, the schemas in (28), repeated below, also help illustrate an assumption I make about the middle-field.

(28) a. [$_{CP}$ XP [V+T+C [$_{TP}$ ~~V+T~~ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]]]]]
b. [$_{CP}$ XP [V+T+C [$_{TP}$ [$_{VP}$ Subj [$_{VP}$ Obj ~~V~~]] ~~V+T~~]]]

I follow Rosengren (2002) in assuming that the relative free word order in the middle-field can be derived by adjoining arguments and adjuncts to VP in the desired order (cf. Haider 2010). I further assume that arguments and adjuncts in the middle-field can adjoin to TP, but in keeping with Rosengren, there is no requirement for the subject to do so. The subject may stay in-situ in SpecVP (28), or it may adjoin to the left of any VP-adjoined adverbs. The significance of this treatment of the middle-field relates to a claim I make later about the children's grammars. In Section 5.3 I discuss how the children might sometimes only be raising the verb to T, even in matrix clauses. Even if this is the case, under Rosengren's analysis of the middle-field it is still possible to have productions in which a post-verbal subject co-occurs (in various positions) with other post-verbal arguments and adjuncts. Indeed, Schönenberger (2001) reports that the Swiss German children do have such productions.

Second, I assume that VP is also head-final. This can be seen by looking at the position of the non-finite verb (which does raise to T) in both matrix (30a) and embedded clauses (30b):

(30) a. De Rochus hät för sini Fründ die Guezli pachet. (=2a)
the Rochus has for his friends these cookies baked
'Rochus has baked these cookies for his friends.'

b. De Prospero isch froh [wenn-er achli in Rue läse cha]. (=4b)
the Prospero is glad when-he a.bit in quiet read can
'Prospero is happy when he can read a bit in peace and quiet.'

I assume that auxiliaries, as in (30a), are just like full lexical verbs in heading a VP. They are thus head-final, taking a VP as a complement, and raise to T.[11] In the case of post-verbal complement clauses, I assume that they have extraposed to some high clause-peripheral position (cf. Stowell 1981).

Turning now to the child grammar, to account for the production errors I assume that the children have correctly learned to raise the finite verb to T but that children initially mis-set the head-direction parameter with respect to TP. In contrast to the target grammar in which TP is head-final, children initially converge on a T-initial grammar. This error need not have any reflex in matrix clauses. If children also raise the verb to C, then matrix clause productions will be entirely ambiguous with respect to TP-headedness. Further, if children only raise the verb to a head-initial TP, so long as only one constituent precedes the finite verb, then matrix clause productions will appear entirely target-consistent. I discuss this latter possibility further in Section 5.3, and simply note here that the analysis does not hinge on this point.

[11] V-finality is not uncontroversial in Germanic syntax. For example, Zwart (1997) proposes that the direct object shifts to a pre-verbal position from a verb-initial VP. This issue appears to be orthogonal to the acquisition question at hand, as I discuss in note 20 below; for concreteness and simplicity I assume the target grammar is V-final.

With a head-initial TP, errors are now predicted in embedded clauses. Note that the identity of H from Schönenberger's analysis in Section 3.4 has now been replaced by T. Accordingly, the error patterns attested in Stage 1 of the children's development (cf. (12) and (15)) are accounted for in exactly the same way as has been discussed. So long as the children's grammar is T-initial in this stage, verb non-finality results from the verb moving past any constituents in or adjoined to the VP. Verb finality is incidental, then, if only a single verb co-occurs with a subject if the children make use of movement of the subject out of the VP into the middle-field, an option that is available in the adult grammar (cf. Diesing 1992). In Section 5.3 I revisit the position of subjects, and we will see that there is reason to think that in both the adult and child grammars stressless subject pronouns are treated differently from other subjects. For the time being, it is sufficient to assume that both kinds of subjects occur in the same positions. Moreover, the more detailed analysis of subjects does not affect the basic analysis of child errors here.

In Stage 2 I assume the children have begun to learn that TP is head-final in the target grammar. Errors begin to decrease as the target T-final value gradually becomes the more dominant setting, with children now moving the finite verb to a clause-final T in embedded clauses more frequently than they move it to a T-initial position.

Having accounted for the child productions, we can now ask questions about the learning process similar to those with respect to Schönenberger's analysis in Section 3.4. How is it that children temporarily mis-learn a T-initial parameter setting when there is no clear evidence for it in either matrix or embedded clauses (cf. (28) and (29))? And if the children have incorrectly learned T-initial, then how do they overcome whatever is responsible for that to arrive at the target grammar? Before illustrating how the model can provide principled answers to these questions in Section 4.3, I first introduce some of the specifics of this particular implementation of the learning model.

4.2 Overview of the model

In this section I present two core components of the model that will allow us to see in the following section how the learning process can address the acquisition puzzle. These components are the parameter space of the learner and the corpus it learns from. Both components involve considerable simplifications compared to what a child learning Swiss German actually faces, but they are sufficiently complex to model a diverse range of linguistic variation. To the extent that these simplifications allow for a first pass at capturing the acquisition data, they can be thought of as a base upon which to ground further extensions of the model. I conclude this section with an overview of how the model receives input and sets parameters, though I postpone more technical discussion of the priors and the update procedure until Section 5.

I begin with the learner's parameter space. These parameters represent the hypotheses that the learner has about the target grammar. The model will explicitly learn

to set 5 syntactic parameters. These are in fact the 5 parameters, examples (23) through (27), which were discussed in the analysis of the adult grammar in Section 4.1. These parameters were chosen because they, minimally, allow the model to learn the core principles of word order in Swiss German. More specifically, these parameters focus on verb placement and are sufficient for accounting for the position of the verb in both the adult grammar (if set correctly), as discussed above, and in the child grammar (if set incorrectly). Thus I abstract away from explicitly learning any parameters concerning the position of arguments. With respect to specifiers, I assume that all occur on the left. Further, having 5 binary parameters puts the model within the range of the number of parameters considered in other multi-parameter learning models (cf. Gibson and Wexler 1994; Sakas and Fodor 2012) and allows for ample exploration of parameter interaction.

There are thus 2 parameters for head-movement and 3 parameters for head-complement directionality. One simplification is to assume minimal syntactic structure along the clausal spine, namely the projections CP, TP, and VP. This is done for expediency so as to keep the number of parameters relatively low, however the logic of the insight of the model (Section 4.3) does not hinge on this. Adding additional projections, for example *v*P, does not pose an obvious problem to the approach here, and I hypothesize that their inclusion would still allow for modeling the development of the Swiss German children.

I assume that all possible combinations of these parameters' values are allowed by UG. With 5 binary parameters, there are thus 32 possible grammars in the hypothesis space of the learner. This assumes that the Head Movement Constraint (cf. Travis 1984) is an operative principle in the model, constraining the space of possible grammars. There are thus no restrictions in the hypothesis space concerning the logically possible combinations of parameter values. This results in a rich hypothesis space containing some grammars that recent work such as Biberauer et al. (2014) claims are not attested cross-linguistically. An example of such an unattested grammar would be one in which TP is head-final and VP is head-initial. The unattested grammars violate a constraint that Biberauer et al. call the Final-over-Final Constraint, or FOFC. For present purposes I will assume that FOFC is not an operative constraint on grammars, and we can note that both the target grammar of Swiss German, and the non-target grammar(s) adopted by the Swiss German children fall within the range of grammars that Biberauer et al. claim are attested. Further, in all the simulations of Swiss German and Korean in Chapters 3 and 4, the model can always learn a grammar within this attested range. However, to the extent that FOFC-violating grammars are hard to learn (perhaps unlearnable) as end-states, an important challenge facing the model is to constrain it from reaching a FOFC-violating end-state. After demonstrating some of the model's successes in the case studies on Swiss German and Korean, in Chapter 4 I discuss some of the difficulties involved in constraining the model. I return to this challenge in Chapter 6 in the context of a broader discussion on learning biases, where I offer a proposal for how the model can address this

challenge.

Another important property of the grammar space in this model is that no complete set of sentences that could be generated by any one of these grammars is a proper subset of those of any other grammar (cf. Safir 1987; Atkinson 2001). Thus any consideration of the 'subset problem' is not applicable here, and as discussed in Chapter 2, I will use the simplified, discriminative implementation of the model.

I next turn to the schematic corpus that the model learns from. The model learns from 14 schematic types of sentences, again making the model comparable to other learning models (e.g. Gibson and Wexler 1994). These are all declarative clauses and provide a robust variety of different word orders that are representative of many of the core patterns in Swiss German/German.[12] These input types are listed in (31). The input varies according to whether there is a direct object, an auxiliary, a clause-initial object (in matrix clauses only), a clause-initial adverbial X (in matrix clauses only), or an embedded clause. I assume that the learner can distinguish nominal arguments (i.e. subjects from objects), perhaps by means of morphological case or general linking principles concerning thematic roles.

(31) *Schematic corpus for Swiss German*

a. SV	e. SV[Comp. SV]	i. XVS	m. OVS
b. SVO	f. SV[Comp. SOV]	j. XVSO	n. OAuxSV
c. SAuxV	g. SV[Comp. SVAux]	k. XAuxSV	
d. SAuxOV	h. SV[Comp. SOVAux]	l. XAuxSOV	

There are two important considerations regarding embedded clauses in the corpus in (31). The first concerns the range of matrix clauses that embedded clauses co-occur with. None of the input types in (31) provides unambiguous evidence for the target setting of T-final, although as was discussed in Section 4.1, certain pairings of embedded and matrix clauses (e.g. *[XVSO [Comp. SOV]]*) do provide such unambiguous evidence. A question, then, is whether including such unambiguous evidence would prevent the model from temporarily mis-setting the TP-headedness parameter to [T-init]. The second point concerns the structural position of embedded clauses, and in particular whether they extrapose. The model currently treats all embedded clauses as being right-adjoined to TP, but one might wonder how a more fine-grained approach to the position of embedded clauses might have an impact on the learning results. I defer further discussion of these

[12] It is interesting to consider how the model might fare if other types of clauses, such as interrogatives and imperatives, both of which are common in child directed speech, were to be included in the corpus. The clearest effect such additional input could have would be to increase the proportion of unambiguous evidence that the verb moves. For example, in polar questions the inflected verb appears clause-initially in the matrix clause. In such a position, it would precede the subject and would unambiguously inform the learner that the verb moved out of the VP. As discussed below regarding examples (33) and (35), in my analysis of the children's errors, V-to-T movement plays a crucial role in both mis-setting and correctly resetting the TP-headedness parameter. With the inclusion of this additional input, then, we still expect modeling results that can capture the Swiss German children's development.

two issues until after having introduced in Section 4.3 the basic insight of how the model can mis-set and reset the TP-headedness parameter. There we will see that in an enriched model that more adequately addresses these issues, we still expect to see the model temporarily learn parameter mis-setting to [T-init], as well as being able to recover and learning the target grammar. This expectation is based on preliminary results from a pilot study, the schematic corpus of which was modified with an eye toward addressing these issues.

The schematic corpus I have proposed in (31) thus provides us with a coarse approximation of Swiss German, but the input is rich enough that the model is still adequately equipped to learn the target grammar given the input that it does receive. It is expected that enriching the schematic corpus would not prevent this. Moreover, based on the preliminary results of the pilot study mentioned above, enriching the corpus is not expected to obviate a learning path that contains the desired non-target stage with a mis-set parameter for [T-init].

An important component of the model is the frequency at which the input types are presented to the learner. Input is sampled randomly and presented to the learner item by item, and the probability distribution over input types that the input is sampled from is based on averages taken from Germanic corpora. This is a simplified way of approximating a subset of the input that we might reasonably expect a learner of Swiss German to be exposed to. The probability distribution of input types is simply a multinomial whose values are arrived at in the following two ways. The probability of an input type could simply be its average frequency rate in a corpus. Or the probability of an input could be the joint probability of certain schematic elements co-occurring; this is calculated by using the probabilities (i.e. the average frequency rates) of the schematic elements in question, assuming they are independent of each other. For example, if half the clauses in a corpus contain an auxiliary and if half are transitive, then the likelihood of a clause that is transitive with an auxiliary is .25. The probabilities of the different input types are given below in (32).

I used the following sources for general guidelines in estimating corpus frequencies; for more details see notes 13-16. First, I follow Yang (2002: 57) in extrapolating the following basic frequencies from an English corpus of child directed speech in CHILDES (MacWhinney 2000) to a V2 grammar: Yang found that roughly half of all sentences are transitive and roughly half contain an auxiliary. Assuming that the occurrence of any schematic vocabulary element is completely independent, this means that the within the groupings of input (32a-d), (32e-h), (32i-l), and (32m-n), each input type is equally probable. Second, the corpus frequency of a subject-initial matrix clause in V2 Germanic is 70% (Lightfoot 1997), and the corpus frequency of an embedded clause in German is approximately 15% (Sakas 2003). As an example, then, the joint probability of having a subject-initial matrix clause and of not having an embedded clause is (.7)(.85) = .595.

The frequencies of the different types of input in (32) can be characterized as

follows. The most common type of input is (32a-d), sentences with an initial subject and no embedding. These occur 59.5% of the time. In contrast, embedded clauses (32e-h) occur only 13.65% of the time, however this input is compatible only with grammars that have a C-initial setting. Input that is compatible only with grammars that have raising of some sort, and are thus [+V-to-T], is robust, occurring 34.56% of the time. Input requiring [+V-to-T] is (32d), as well as the input with a post-verbal subject in (32i-j) and (32l-m). Similarly, input with an auxiliary and an object, which is compatible only with V-final grammars, occurs at a rate of approximately 25.5%. Finally, no input is compatible only with either T-initial or T-final, and the same holds for [±T-to-C] movement. Nevertheless, only a [+TC, T-fin] grammar is compatible with all the input.

(32) *Input type and probability* [13]

a. SV	$p = .14875$		e. SV[Comp. SV] [14]	$p = .034125$	
b. SVO	$p = .14875$	.595	f. SV[Comp. SOV]	$p = .034125$	.1365
c. SAuxV	$p = .14875$		g. SV[Comp. SVAux]	$p = .034125$	
d. SAuxOV	$p = .14875$		h. SV[Comp. SOVAux]	$p = .034125$	

i. XVS [15]	$p = .06265$		m. OVS [16]	$p = .00895$	.0179
j. XVSO	$p = .06265$	.2506	n. OAuxSV	$p = .00895$	
k. XAuxSV	$p = .06265$				
l. XAuxSOV	$p = .06265$				

I have just noted that around one-third of the corpus in (32) unambiguously provides evidence for verb movement. Bobaljik (2000, 2001) raise an interesting possibility that inflectional morphology can inform the learner about V-to-T movement. According to Bobaljik's version of the Rich Agreement Hypothesis and his theory of feature checking, if there are at least two pieces of verbal morphology (each with its own syntactic projection) on the finite verb, then the verb must raise to T (see Bobaljik for references for, and arguments against, other formulations of the Rich Agreement Hypothesis). No such morphological detail is provided in the corpus in (32), and one

[13] Note that because not all possible combinations of schematic elements occur in the schematic corpus, I adjusted the probabilities of the input types so that they sum to 1. Without any adjustment, .045 of the probability mass was unaccounted for. This corresponds to the joint probability of a non-subject-initial matrix clause co-occurring with an embedded clause: (.3)(.15) = .045; see notes below for the sources of these probabilities. (None of the non-subject-initial matrix clauses co-occur in the corpus in (32) with embedded clauses.) Accordingly, I decided to spread this .045 probability mass out across the 12 kinds of input that were either non-subject-initial or contained some embedding (32a-d). The probability mass of .045 was redistributed based on the frequency of these kinds of input in the corpora. For example, input with X occurs in the corpora 28% of the time; thus (.045)(.28) = .0126 was added to .238 (the probability of X input not co-occurring with an embedded clause), giving a probability mass of .2506, which was then divided evenly across the four kinds of input with an initial adverbial X. This adjustment amounts to a relatively minor smoothing out of the averages for input frequency in the corpus. Given that these averages are based in part on corpora of child directed speech, there is reason to think they provide the model with a solid grounding. Still, the averages themselves are coarse approximations of the child's input, and given that the adjustment concerns a small proportion of the probability mass, I do not expect much to hinge on the adjustment itself.

[14] Sakas (2003) reports that on average 15% of a child's input contains some embedding. This is based on an analysis of several thousand sentences from corpora of 5 languages (including German) in CHILDES (MacWhinney 2000). The joint probability of having an embedded clause and of having that clause be introduced by a subject-initial matrix clause is (.7)(.15) = .105. This probability mass was adjusted to .1365 as discussed in note 13.

[15] The corpus frequency of a clause-initial constituent other than the subject or direct object in matrix clauses in V2 Germanic is 28% (Lightfoot 1997). The joint probability of having a matrix clause with such an initial constituent and of that matrix clause not introducing an embedded clause = (.28)(.85) = .238. This probability mass was adjusted to .2506 as discussed in note 13.

[16] The corpus frequency of a clause-initial direct object in matrix clauses in V2 Germanic is 2% (Lightfoot 1997). The joint probability of having a matrix clause with such an initial direct object and of that matrix clause not introducing an embedded clause = (.02)(.85) = .017. This probability mass was adjusted to .0179 as discussed in note 13.

might wonder whether important evidence concerning the position of the verb is missing in the corpus. Although such detail could provide a helpful clue to the learner, the general approach I am advocating is that to learn the presence or absence of verb movement it is sufficient for the learner to rely on word order in conjunction with parameter interaction.[17] Moreover, there is some evidence from Korean that runs counter to Bobaljik's theory. In complex verbal predicates in Korean, it is certainly plausible to treat the multiple suffixes on the verb as instantiating different pieces of inflectional material (for an example, see Section 3.2 in Chapter 4). Nevertheless, a central claim that I make is that some learners of Korean do adopt a grammar that has no verb movement. Bobaljik's theory is thus neither necessary for the acquisition cases at hand and is perhaps too restrictive cross-linguistically.

Having discussed the model's parameters and input, I now present an overview of how the model learns from the input. This follows the general framework of the model presented in Chapter 2. More precise discussion of the priors and update procedure can be found in Section 5.

In particular, I will be using the simplified implementation of the model introduced in Chapter 2. Upon encountering a token of input, the model samples a parameter value from the probability distribution of each parameter to form a possible grammar, or vector of parameter values. Initially, all parameter values are assumed to have equal priors. The sampled values are then compared with the input. If the sampled grammar is compatible with the input, then the probabilistic weights of those values are increased. This is done so as to maximize the likelihood of sampling such a compatible grammar in the future given the same input. In this respect, the model is similar to Yang's learning model (cf. Section 6). In terms of the fully generative model from Chapter 2 (which outputs sentences that are compared against the input), if output matching the input could possibly be generated with the parameter values sampled, then those values are reinforced.

Unlike the fully generative version of the model, the simplified implementation here only checks compatibility between sampled parameter values and the input. I thus abstract away from a number of syntactic details that would be involved in generating output strings that match the strings in the schematic corpus. For example, the 5 parameters in the model do not take into consideration how modifiers are represented in the syntax; a fully generative model that outputs full strings would need to take this into consideration.[18] As discussed in Chapter 2, this simplifying step allows for a more

[17] Thus input with inflected verbs is often ambiguous with respect to V-to-T movement. To take a simple example, SV input with an inflected verb is compatible with either a [+VT] or [–VT] grammar. If the verb remains in-situ, I assume that inflectional material can affix to the verb either via some rule of affix lowering (e.g. Chomsky 1957) or some post-syntactic process such as Morphological Merger (Marantz 1988); cf. Chapter 4 for additional discussion of verbal affixes.

[18] I also abstract away from parameters concerning the position of nominal arguments, in particular the subject. Given the relative flexibility in ordering elements in the middle-field, formulating such parameters is not trivial, and I leave it as a task for future research. Preliminary analysis points to this flexibility

expedient proof-of-concept illustration of how the model can account for errors. A fully generative model will be crucial in Chapter 5 when learning from implicit negative evidence in the case of languages are in a subset/superset relation with each other. Recall that I assume that no complete set of sentences that could be generated by any one of the 32 grammars in the hypothesis space here is a proper subset of those of any other of the 32 grammars. The only differences across the grammars, then, are the settings for the five parameters under consideration. All things being equal, adding an enriched probabilistic context-free grammar, with further probabilistic choices that do not interact with these five parameters, should not affect the overall learning course of the model presented here. Thus it is possible to use the simplified version of the model from Chapter 2.

Let us take a look at an example of an input type to see what kinds of parameter vectors are compatible with it. A full list of all the grammars that are compatible with each type of input can be found in Appendix 1 at the end of this chapter. As an example, let us consider SV input, which is compatible with all 32 possible grammars. As there are no parameters in the model that explicitly concern the position of the subject, the model accounts for the subject 'for free' as it were. This means that all other parameter settings can be consistent with SV input. To see why this is so involves considering how high the verb has raised in the structure. If the verb does not raise, staying in-situ in the VP, the subject can appear in any leftward specifier: SpecVP, SpeccTP, or SpecCP (perhaps in case of topicalization). As no other constituent appears within the VP, either V-final or V-initial is a possible setting (assuming some kind of affix lowering for inflectional morphology). And as the verb has raised to neither T nor C, either one could be head-initial or head-final. If the verb raises to T, then the subject can no longer remain in the VP: in order for the subject to appear-preverbally, it must move to SpecTP or SpecCP. However, as nothing follows the verb, either final or initial are possible settings for T and V; and as before when the verb is not in C, C-initial or C-final are possible. Finally, if the verb raises all the way to C, then the subject must be in SpecCP. But nothing else about head-complement directionality is determined because there are no post-verbal constituents.

Thus we see that SV input is fully ambiguous for all 32 possible combinations of parameter values. Indeed as shown in Appendix 1, nearly all input types are compatible with more than one grammar. But not all types of input are compatible with all grammars. In fact, some types of input are compatible with more grammars with one value of a parameter than with that parameter's opposite value. In particular, a large proportion of the input favors a setting of T-initial over T-final, a point which I explore in more detail below.

---

introducing a great deal of ambiguity regarding the position of the subject. The upshot of this ambiguity appears to be that ambiguous matrix clauses still favor a non-target setting of [T-init] as the most probable analysis via parameter interaction (cf. Section 4.3 for how the model learns this non-target value). Thus even with the inclusion of further parameters concerning the position of the subjects, the outlook is promising for capturing the children's [T-init] parameter mis-setting.

4.3 Insight of the model

In this section I show that a prediction of the model is that due to parameter interaction, a Swiss German learner will on average initially favor a T-initial setting. According to my analysis, child production errors are the result of using such a non-target grammar, i.e. raising the verb to an initial T. By carefully examining which grammars are compatible with which parameter settings, we can see how a learner can be pushed initially to favor a T-initial setting. I illustrate this with one core type of input: SVO declaratives. An important consideration is that all the input underdetermines the correct structural analysis of TP-headedness. We have already seen this with matrix clauses in (23), which are ambiguous for T-final/initial. This is also true for embedded clauses, a point I return to below. There is thus no input that unambiguously 'flags' to the learner that the target grammar is T-final (i.e. there is no global trigger in the terms of Gibson and Wexler 1994). To see the role this ambiguous input plays in the learning process, it is necessary to look at all the grammars that are compatible with a given input to see whether one value for TP-headedness is favored over the other.

We saw at the end of the previous section that all of the 32 logically possible grammars are compatible SV input; half of these are T-initial, and half are T-final. However, there are 16 grammars that are compatible with SVO input, listed below, but now a majority of these are T-initial grammars.

(33) *Grammars compatible with SVO input*

a. [+VT, +TC, C-init, T-init, V-init]
b. [+VT, +TC, C-init, T-init, V-fin]
c. [+VT, +TC, C-init, T-fin, V-init]
d. [+VT, +TC, C-init, T-fin, V-fin]

e. [+VT, –TC, C-init, T-init, V-init]
f. [+VT, –TC, C-init, T-init, V-fin]
g. [+VT, –TC, C-fin, T-init, V-fin]
h. [+VT, –TC, C-fin, T-init, V-init]

i. [–VT, –TC, C-init, T-init, V-init]
j. [–VT, –TC, C-init, T-fin, V-init]
k. [–VT, –TC, C-fin, T-fin, V-init]
l. [–VT, –TC, C-fin, T-init, V-init]

m. [–VT, +TC, C-init, T-init, V-init] [19]
n. [–VT, +TC, C-init, T-fin, V-init]
o. [–VT, +TC, C-fin, T-fin, V-init]
p. [–VT, +TC, C-fin, T-init, V-init]

[19] A [–VT, +TC] language could involve movement of inflectional affixes from T to C, with the verb remaining in-situ. This is reminiscent of Zwart's (1997) analysis of complementizer agreement in Germanic, but a full treatment of that phenomena under the learning model goes beyond the scope of this study.

10 of these grammars are T-initial, whereas 6 of them are T-final. Crucially, we see that verb-movement to T interacts with TP-headedness: a [+VT, –TC] grammar requires T to be initial. In a [+VT] grammar, the subject can precede the verb by being in SpecTP or SpecCP, but because the object is post-verbal, T cannot be final.[20] All things being equal, the model predicts that a grammar with a T-initial value is more likely to be sampled by the learner when encountering SVO input. A T-initial grammar is more compatible given such input. And by sampling a T-initial value, the probabilistic weight for that value will be increased, thereby increasing the likelihood that a T-initial value will be sampled again. The learning process, supported by this ambiguous input, reinforces the selection of a T-initial value, pushing the learner toward a greater and greater weighting for a T-initial setting.

Going beyond SVO input, the same parameter interaction results in more T-initial grammars being compatible with nearly all kinds of input. Just as with SVO input, in other kinds of input (e.g. XVS, SAuxV, etc.) with a post-verbal constituent that occurs below T, a compatible [+VT, –TC] grammar must have an initial T. With parameter settings of [+VT, +TC], [–VT, –TC], or [–VT, +TC], T can be final or initial. The balance of grammars, then, will always be in favor of T-initial for these kinds of inputs.

Of the 14 types of input, 9 types favor T-initial grammars. These are 9 out of the 10 types of matrix clauses: (a) matrix clauses with a post-verbal subject (assumed to be in SpecVP) because of an initial adverbial or object; and (b) matrix clauses with a clause-final non-finite verb, which co-occurs with an auxiliary. The two types of clauses that do not favor T-initial are (a) matrix and embedded SV input, which favor neither T-initial nor T-final; and (b) all other embedded clauses, which favor T-final. This can be verified by consulting the table in Appendix 1. A classification of these clause types is given below.

[20] I note that a V-initial analysis of German along the lines of Zwart (1997), in which the object shifts to a pre-verbal position from a verb-initial VP, has no bearing on T-initial being favored over T-final. So long as the object appears post-verbally below T, a T-initial setting occurs in the majority of compatible grammars. For concreteness and simplicity I assume that the target grammar is V-final, though this does not appear to be crucial for the learning model.

(34) a. *Input favoring T-initial*

1. SVO	4. XVS	8. OVS
2. SAuxV	5. XVSO	9. OAuxSV
3. SAuxOV	6. XAuxSV	
	7. XAuxSOV	

b. *Input favoring T-final*
1. SV[Comp. SOV]
2. SV[Comp. VAux]
3. SV[Comp. SOVAux]

c. *Input favoring neither T-initial nor T-final*
1. SV 2. SV[Comp. SV]

Figure 3.1 shows the extent to which T-initial is favored in the corpus. The columns in the chart represent different sets of input types. These are grouped together based on having similar proportions (as measured on the Y-axis) of T-initial grammars within the sets of grammars that are compatible with them. The widths of these columns correspond to the percentage of the corpus represented by the different groupings of input (based on their probabilistic frequencies in (32)).

What we see in Figure 3.1 below is that out of an entire corpus that is ambiguous for TP-headedness, 71.745% of it favors a T-initial grammar, whereas only approximately 10.24% of the input favors a T-final grammar. Moreover, the input that favors T-initial strongly favors it with proportions ranging from .625 to .75. And the more strongly a parameter value is favored overall, the more likely a learner will be pushed toward a stronger weight for that value and thus toward adopting that value as a parameter setting. In an input-driven learning model, then, the prediction is that on average the learner will initially be pushed toward a T-initial grammar.

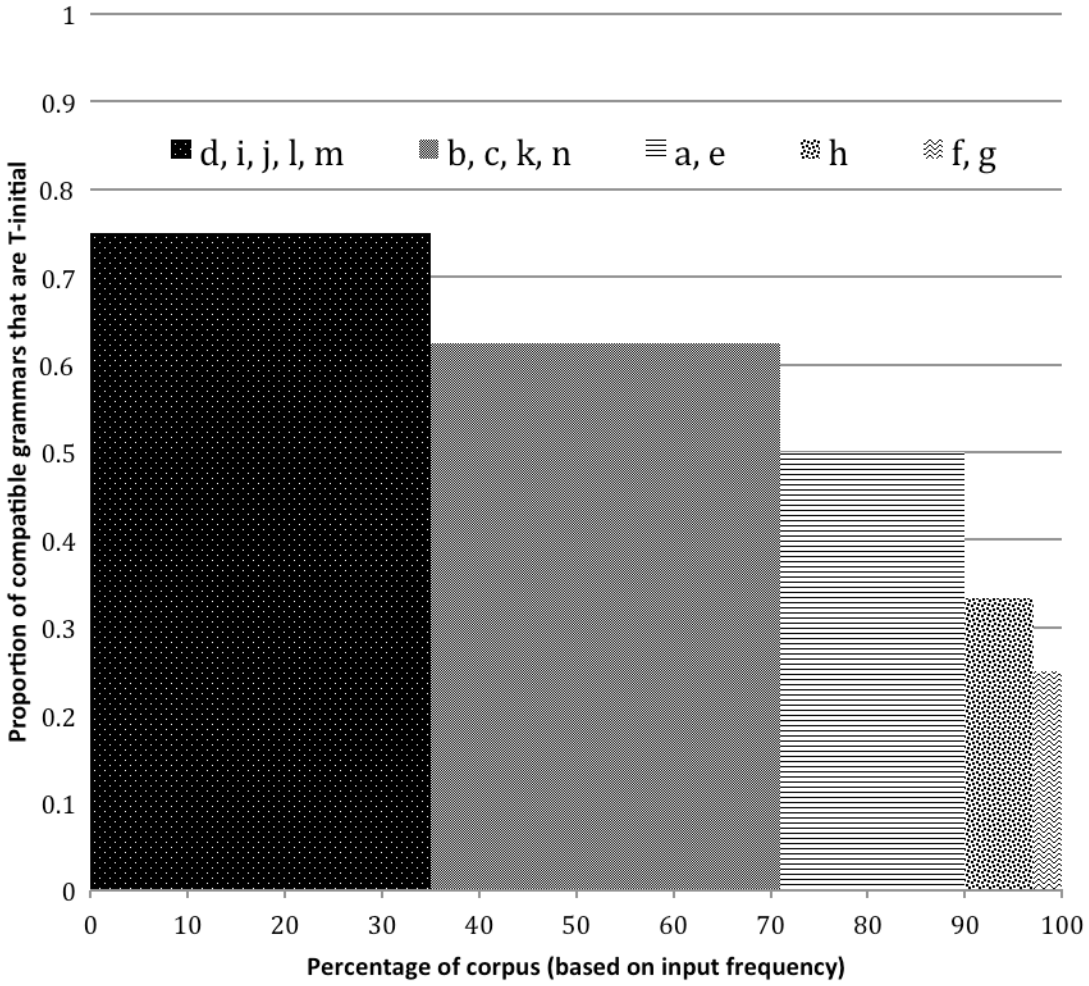
Given that parameter interaction favors T-initial rather strongly, and in line with my analysis of the children's grammar, the hypothesis I will pursue is that this parameter interaction could, at least sometimes, push the learner all the way to mis-setting the TP-headedness parameter to [T-init]. Importantly, this result would fall out from the nature of the input and the learning process itself. We now have a principled account of why a learner would adopt a grammar heavily weighted toward T-initial.[21]

[21] Having seen how parameter interaction can systematically push the learner toward T-initial in the 5-parameter hypothesis space, we can return to the 6-parameter hypothesis space that I mentioned in note 10. This hypothesis space includes an additional parameter for V-to-C movement, [±VC]. How well might the V-to-C hypothesis space fare in accounting for learning errors? In brief, the 6-parameter hypothesis space is less likely to have results that model the errors of the Swiss German children than the 5-parameter hypothesis space is. To see this, let us consider how the children's errors would arise in the V-to-C hypothesis space.

In the V-to-C hypothesis space, there is no longer unambiguous evidence for [+V-to-T] because whenever the verb appears outside of the VP, it might have done so under either a [+VT] or a [+VC, –VT] grammar. Still, analytical investigation reveals that [+VT] is favored via parameter interaction, and so we expect learners to be pushed toward a [+VT] grammar. If the model learns a [+VT] parameter setting, then the model would also need to learn T-final to account for the input (for reasons similar to those discussed in the main text), and the children's errors could then be accounted for by parameter mis-setting to [T-init] as discussed in the main text. Thus the grammar of best fit is T-final, and given sufficient input we expect the model to learn such a grammar.

The question is whether parameter interaction in the V-to-C hypothesis space can push the learner sufficiently toward [T-init] in order to model the near-ceiling error rates we see in the Swiss German children. I have not given any precise criterion for how strong parameter interaction must be in order to result in parameter mis-setting, and so I must be rather speculative here. What we can observe is that parameter mis-setting is less likely with a V-to-C hypothesis space. The reason for this lower likelihood is as follows. First, the distribution of input types in (34) is the same in the V-to-C hypothesis space. However, what is different is that the degree to which the input favors [T-init] has been greatly attenuated. In the V-to-C hypothesis space, the proportion of input-compatible grammars that are T-initial never exceeds 59%. This is noticeably less than the degree to which T-initial is favored in the 5-parameter hypothesis space: as shown in Figure 3.1, the proportion of input-compatible grammars that are T-initial always exceeds 62% in the 5-parameter hypothesis space. The reason for lower proportions in the V-to-C hypothesis space is because the number of grammars in which the verb is in C has increased. This in turn lowers the proportion of grammars in which the verb moves only as high as T; as we saw in (33) with SVO input, it is these grammars that provide an advantage to learning T-initial. Thus on average, we expect the learner to be less likely to be pushed toward a stronger weight for T-initial under the V-to-C hypothesis space than under the 5-parameter hypothesis space. Consequently, parameter mis-setting is less likely with the V-to-C hypothesis space. The results of the simulation of the 5-parameter space in Section 5 show that the degree to which T-initial is favored in Figure 3.1 is sufficient to result in parameter mis-setting. In sum, although with the V-to-C hypothesis space we would expect the learner to on average be pushed toward a T-initial grammar initially, it remains to be seen whether the learner would be sufficiently pushed toward parameter mis-setting of [T-init]. Without running further simulations of the model with the V-to-C hypothesis space, it is difficult to conclude anything further about the inclusion of a V-to-C parameter, and I must leave this question for future research.

Figure 3.1 Input type by frequency and strength of favoring T-initial



d. SAuxOV *b. SVO* *a. SV* *h. SV[Comp. SOVAux]* *f. SV[Comp. SOV]*
i. XVS *c. SAuxV* *e. SV[Comp. SV]* *g. SV[Comp. SVAux]*
j. XVSO *k. XAuxSV*
l. XAuxSOV *n. OAuxSV*
m. OVS

However, an important question arises about the role of embedded clauses in the learning process. If embedded clauses are crucial in learning that the target grammar is T-final, and since the learner receives input of embedded clauses from the very beginning of the learning process, one could ask why the ambiguous matrix clauses would have much of an effect on the learner.

There are two related factors in mis-setting the TP-headedness parameter. First, the effect of embedded clause input is initially mitigated by the fact these clauses are also ambiguous with respect to TP-headedness. To take an example, even though SV[Complementizer SOV] input favors T-final, T-initial is still compatible so long as [–VT, V-fin] is chosen:

(35) *Grammars compatible with input SV[Complementizer SOV]*

a. [+VT, +TC, C-init, **T-fin**, V-init]
b. [+VT, +TC, C-init, **T-fin**, V-fin]

c. [+VT, –TC, C-init, **T-fin**, V-init]
d. [+VT, –TC, C-init, **T-fin**, V-fin]

e. [–VT, –TC, C-init, T-init, V-fin]
f. [–VT, –TC, C-init, **T-fin**, V-fin]

g. [–VT, +TC, C-init, T-init, V-fin]
h. [–VT, +TC, C-init, **T-fin**, V-fin]

Thus the learner cannot simply rely on embedded clauses as unambiguous evidence for T-finality. Initially, the informativeness of embedded clause input with respect to T-finality is greatly diminished. Second, the edge that embedded clauses can give initially to T-final in (35), for example, can be outweighed by matrix clause input that overwhelmingly favors T-initial. The vast preponderance of all input is matrix clauses that favor T-initial. This is especially true given the randomness of the input, such that the learner can sometimes be exposed to stretches of input in which there are hardly any embedded clauses. Again, the informativeness of embedded clauses initially can be quite small.

Given these factors, a related question is the following: supposing that matrix clauses do have a strong pull on the learner toward a T-initial grammar, how then does the learner converge ultimately on the target T-final grammar? Importantly, the learner must receive a stretch of input with a sufficient amount of embedded clauses, as it is embedded clauses that ultimately provide the evidence for parameter resetting. Indeed, given the randomness of the input, at some point the learner will comes across this evidence. There are also several additional factors that play a role here, and they are to a certain extent inter-related in pulling the learner toward the target grammar.

First, learning [+VT] can help the model converge on the target grammar. Learning [+VT] is straightforward enough because there is unambiguous evidence for that setting in matrix clauses: in addition to SAuxOV strings, the input with post-verbal subjects in (32i-j and l-m) is only compatible with [+VT] grammars. Further, because unambiguous evidence for [+VT] is so robust (34.56% of the corpus), on average we might expect weighting in favor of [+VT] to be heavier (with a probability approaching 1) than weighting in favor of T-initial. An embedded clause under a [+VT] setting must be T-final, and if a [+VT] weighting is heavier than a [T-init] weighting, then the T-initial value will then begin to lose probabilistic weight to the T-final value. Given a setting of [+VT], the shift from T-initial to T-final is driven by a stretch of input with a sufficient amount of embedded clauses in the input, (i.e. embedded clauses that are then only compatible with T-final). As these clauses comprise only 10.24% of the input, we expect the shift to T-final to be a gradual one. This is in fact what we see in Stage 2 of the

children's development, during which the occurrence of clause-final finite embedded verbs becomes more and more prevalent.

Second, it can also happen that in mis-setting [T-init], the weighting for [T-init] is as heavy as that of [+VT]. For this to happen, the model would have needed to have had a high sampling rate for [T-init], but this can happen, especially with an abundance of matrix clause input that heavily favors [T-init]. Still, there are more [+VT] grammars, (35a-d), all of which are [T-fin], than there are [T-init] grammars, (35e, g), that are compatible with embedded clauses. The additional [+VT] grammars (35a, c) are both [V-init]. Suppose the model has come close to learning [V-fin], but still has a not too insignificant weighting for [V-init]. This would mean that the model has weightings for [+VT] and [T-init] that are stronger than the weighting for [V-fin]. This possibility can arise, since there is more unambiguous evidence to [+VT], around one-third of the input, than there is unambiguous evidence for [V-fin], which is around one-fourth of the input. This extra bit of probability mass for [V-init] can help push the learner toward a [+VT, T-fin] grammar. Again we see that, a heavy weighting for [+VT], this time in conjunction, with a relatively light weighting for [V-init], will have more influence in shifting the grammar than a heavy weighting for [T-init] will.[22]

There is a third aspect to parameter resetting. Suppose that [+VT] and [T-init] have equally heavy weightings, and that the learner has set the VP-headedness parameter as [V-fin]. Now there might not appear to be anything to choose between, say, the [+VT, T-fin] grammar in (35b) and the [–VT, T-init] grammar in (35g). Both might appear to be equally good. Still, when the model samples for a compatible grammar, something has to give: either the learner retreats from [T-init] or abandons [+VT]. Ultimately for any given sample, the [T-fin] grammar is the more likely choice because there are simply more [T-fin] grammars in (35), for example, than there are [T-init] grammars. In the long run, a stretch of input with a sufficient amount of embedded clauses can thus push the learner toward a [T-fin] grammar. As the learner thus retreats from [T-init], [+VT] will have the stronger weighting than [T-init], and we return to the first factor from above, in which a strong weighting for [+VT] pushes the learner toward a [T-fin] analysis of embedded clauses.

All of these factors can help the learner reset the parameter value to [T-fin], given a sufficient amount of embedded clauses. Thus we see that learning [+VT] has a conditioning effect on learning [T-fin]. This falls under the rubric of what I call *secondary parameter interaction* in Chapter 4 (Section 5). For expository purposes, I reserve a more formal discussion of this phenomena for Chapter 4 where we will see a subtler effect of this kind of interaction on the degree of grammatical variability within a

[22] One might wonder whether this extra bit of probability mass for [V-init] during Stage 1 results in the model sometimes adopting a non-target [V-init] grammar at the time of mis-setting [T-init]. In the simulations of the model reported in Section 5, the probability for [V-init] is small enough that this occurs on average less than 5% of the time (cf. the first cell under 'Other Grammars' in Table 3.2 in the results below).

population of learners.

In summary, embedded clauses always remain a relatively small proportion of the learner's input. Given the randomness of the input, embedded clauses will sometimes be very infrequent indeed, and thus there is always the possibility for stretches of learning in which embedded clauses play very little role in the learning process. I have discussed how parameter interaction can push the learner toward parameter mis-setting. In conjunction with parameter interaction, such stretches of input with a low frequency of embedded clauses can also contribute to the learner mis-setting a parameter early on. The difference between parameter mis-setting and parameter resetting involves two separate factors. First, with a strong weighting for [+VT], embedded clauses are on average more likely to be T-final, and this increase in likelihood helps reset the parameter. Second, given the randomness of the input, at some point after learning [+VT] the learner will then be exposed to a stretch of input with a sufficient amount of embedded clauses to recover from the non-target grammar.[23]

This is not to say that in mis-setting [T-init], learning [+VT] is delayed; rather, there is enough of a window between the initial state and the learner's heavily favoring [+VT] for the learner to also simultaneously be pushed strongly toward [T-init]. Indeed, the process of learning [+VT] helps push the learner toward [T-init], as was discussed regarding (33) above. But once a very strong weight has been assigned to [+VT], then on average, a [T-fin] analysis of embedded clauses becomes much more likely.

So far in this section I have discussed how parameter interaction, input frequencies, and how the model learns can push the learner toward parameter mis-setting, while still being able to recover and reset the parameter to the target value. In the remainder of this section I present a more detailed discussion of two points relating to embedded clauses in the input, a discussion that I have deferred up to this point. These two issues concerned unambiguous evidence for T-final, as well as the structural position of embedded clauses. With the insight of how parameter interaction works in hand, as well as preliminary results from a pilot study that I introduce below, we will see that an enriched model with a more fine-grained treatment of embedded clauses do not pose any obvious difficulties in modeling parameter mis-setting and resetting.

First, I consider input types containing embedded clauses that provide unambiguous evidence for T-final. Recall from the schematic corpus in (31), repeated below, that embedded clauses co-occur only with S and V in matrix clauses. This is a simplification, and its primary purpose was to keep the schematic corpus compact while still maintaining

[23] The discussion here raises interesting questions about the frequency of embedded clauses both at the time of parameter mis-setting and parameter resetting. For example, to what extent must this frequency be relatively low at the time of mis-setting, and to what extent must this frequency be relatively high at the time when the learner begins to reset the parameter? Unfortunately I cannot address these questions here because the simulations of the model reported in Section 5 do not contain information on the shape of the input beyond the corpus frequencies in (32). Nevertheless, in both parameter mis-setting and resetting we can point to parameter interaction as playing an important role, as has been made clear in the discussion above.

a diversity of input types. Further, we will see evidence from a pilot study that this simplification does not call into question the basic results of the simulation with the corpus in (31).

(31) *Schematic corpus for Swiss German*

a. SV	e. SV[Comp. SV]	i. XVS	m. OVS
b. SVO	f. SV[Comp. SOV]	j. XVSO	n. OAuxSV
c. SAuxV	g. SV[Comp. SVAux]	k. XAuxSV	
d. SAuxOV	h. SV[Comp. SOVAux]	l. XAuxSOV	

The simplification pertains to the amount of unambiguous evidence for [T-fin] and [+T-to-C]. Recall from the discussion in the previous section that some embedded clauses provide unambiguous evidence for these parameter values when they co-occur with certain matrix clauses. Specifically, these are sentences that combine the embedded clauses with an object or an in-situ verb in (31f-h) with any of the matrix clauses in (31b-d) and (31i-n). None of these types of input occur in the corpus above. Consequently, there is no unambiguous evidence for [T-fin] or [+T-to-C]. This does not pose a problem, though, for learning the target grammar, which has those values. Only a grammar with those values is compatible with every type, and it is such a grammar that will emerge as the grammar of best fit in the learning model.

However, the absence of unambiguous input for [T-fin] raises a question regarding parameter mis-setting. Under the analysis I have proposed, the model first mis-learns a parameter value of [T-init]. In this section we have seen how the mechanics of the model allow this to be possible because of parameter interaction that at first favors [T-init] (while being able to recover and reset the value to [T-fin]), and in Section 5 we will see results of a simulation in which such parameter mis-setting does occur. These results are based on learning from the corpus in (31) in which there is no unambiguous evidence for [T-fin]. One might wonder, then, whether the presence of such unambiguous evidence might prevent the model from mis-learning [T-init].

Preliminary results from a pilot study indicate that even if unambiguous evidence were included, the model can still mis-learn [T-init]. The pilot study is minimally different in that the embedded clauses in (31f-h) were treated as if they were unambiguous for [+T-to-C, T-fin]. Additionally these embedded clauses occur at the same frequency as in the current study (cf. (32) above). This was to simulate these embedded clauses co-occurring with non-SV matrix clauses, the combination of which provides unambiguous evidence for the two parameters in question. With this new source of unambiguous evidence, the new simulation is able to move toward the adult grammar more quickly, but there is still a period of time in which a non-target [T-init] is adopted by the learner; this corresponds to Stage 1 of the Swiss German children's development. Using such a non-target grammar to model the children's utterances (cf. Section 4.4), the pilot study currently predicts an error rate as high as 87% for verb placement in

embedded clauses, even after having encountered the unambiguous embedded clauses in the input. This is comparable to the error rate of 94% in the current version of the model, as reported in Table 3.2 below. Thus the pilot study has some early success in replicating the results reported here. Moreover, it does so while likely overestimating the frequency of embedded clauses that are unambiguously [T-fin]. This is because the corpus still does not include embedded clauses that are V2, which are ambiguous as to the position of T with respect to its complement.

Including this unambiguous evidence, then, does not appear to prevent the learner from early parameter mis-setting. This result is not surprising for the following two reasons. In the first place, embedded clauses are not especially frequent in the input. In the input frequencies reported in (32) above we saw that embedded clauses occur in approximately 13% of all input, and the unambiguous embedded clauses comprise only three-quarters of this, or 9.75% of the entire learner's input. Most of the learner's input is made up of ambiguous matrix clauses, and as I discussed in this section, the preponderance of this ambiguous evidence can push the learner toward a [T-init] analysis via parameter interaction. Thus even if the learner is exposed to *some* unambiguous evidence, it is not always sufficient early in the learning process to preclude parameter mis-setting. Second, even though the embedded clauses in (31f-h) in the current simulation are ambiguous for TP-headedness, they still favor a [T-fin] analysis. This was illustrated in (35) above, where it was shown that [T-fin] is the most likely analysis of these embedded clauses because of parameter interaction. Thus the current study is not as different from the pilot study as it may appear. It is really a question of degree, with the unambiguously embedded clauses in the pilot study more strongly pushing the learner toward [T-fin]. This is why the learner more quickly adopts the target grammar in the pilot study.

To reiterate, although the simplified corpus in (31) does not present the learner with the full range of unambiguous evidence available to the Swiss German child, we do not expect this simplification to greatly impact the result of the learning model in which there is parameter mis-setting.

The second issue regarding the embedded clauses in (31e-h) concerns the possibility of complement clauses and extraposition. The embedded clauses have been treated uniformly in (31e-h) as following an SV sub-string. This is certainly not the only position embedded clauses appear in the adult grammar, but moreover the level of detail in (31) neutralizes the difference between adjunct embedded clauses and complement clauses. This point bears some additional discussion with respect to how these data play a role in setting the VP-headedness and TP-headedness parameters. If the embedded clauses in (31e-h) are adjunct clauses, then they could right-adjoin to VP or TP, and the input would be ambiguous with respect to VP-headedness or TP-headedness. However, if the embedded clauses are complement clauses, then as arguments of the verb, they could appear to be sisters of V, and these data could be taken as evidence for a V-initial setting.

This evidence would still be ambiguous: (31e-h) would have to be V-initial only if the verb does not move at all. Still, as I am assuming that the target grammar is V-final, such evidence might provide an additional challenge for the learner. Similarly, if the embedded clause were treated as a sister to V, it would be like an object in SVO strings with respect to TP-headedness. If the verb moved only to T, T would have to be head-initial (cf. (33) above). Again, this could make it harder for the model to learn a T-final setting.

As I noted in Section 4.2, the simulations reported in Section 5 simply treat all embedded clauses as being like adjunct clauses that are right-adjoined to TP. Thus embedded clauses are uninformative with respect to VP-headedness and TP-headedness in matrix clauses. If we follow Stowell's (1981) analysis that complement clauses extrapose and adjoin to some peripheral position, then the learner could treat all complements of the verb, whether clausal or not, as preceding V in their base position. The model does not currently contain any parameters for extraposition, and one might wonder whether input with complement clauses of the form (31e-h) could push the learner toward a non-target V-initial setting.[24] Nevertheless, even without learning extraposition, under a more enriched version of the model that distinguishes complement clauses from adjunct clauses, complement clauses of the form (31e-h) do not obviously pose a substantial obstacle to learning V-final for the learner. First, as we saw in the corpus frequencies in (32) above, embedded clauses form a relatively small proportion of the corpus. Sentence-final complement clauses are only a subset of embedded clauses in Swiss German, and their overall effect on the learner could be minimal. In contrast, input that is unambiguous evidence for V-final, such as (31d, l) is relatively much more common in the corpus. Moreover, complement clauses only favor V-initial via parameter interaction if there is no verb movement. If there is no verb movement, then (31e-h) must be V-initial; if there is verb movement, then either a V-initial or V-final grammar is possible. I have already discussed that a large proportion of the corpus provides unambiguous evidence for verb movement. Indeed, the model rapidly learns that there is verb movement, and this success with learning verb movement would neutralize any advantage that a non-target V-initial setting could gain from complement clauses. Thus it is not expected that distinguishing complement clauses from adjunct clauses in (31e-h) would have much of an effect on the learning trajectory of the model.

Similarly, we do not expect complement clauses to play much of a role in setting the TP-headedness parameter. Recall that if the verb raises to T with the complement

[24] I leave for future research the full ramifications of including extraposition parameters and their interaction with additional types of input, such as clauses with phonologically heavy DP objects that have been extraposed, or utterances with a matrix auxiliary and a complement clause (e.g. [SAuxVCP]), which would be unambiguously V-initial without extraposition. However, the possibility of extraposition means that utterances of the latter type area ambiguous with respect to V-initial (with or without CP extraposition) or V-final (with CP extraposition). Again, such ambiguity attenuates the influence such input can have on favoring a V-initial grammar. Moreover, if utterances of the latter type are sufficiently infrequent, then we do not expect much difference in the learning outcomes.

clause not extraposing, then TP must be head-initial. This can provide a slight edge to the non-target value of [T-init], although it would do so in only the small proportion of the input that has complement clauses. (And its effect would also be attenuated if the learner ever chose to extrapose the clause.) This slight favoring of [T-init] could help in the desired mis-setting [T-init], but it is unnecessary for parameter mis-setting. In the current simulation, which does not distinguish complement clauses, the model is already capable of mis-setting [T-init]. But might slightly favoring [T-init] prevent the model from reaching the target state? Again, this is not expected. The most compatible grammar is still one that is [T-fin], as only such a grammar is compatible with all the input. Further, including the embedded clauses that are unambiguous for [T-fin] in the pilot study mentioned above will further reduce the frequency of complement clauses that favor [T-init]. In the pilot study such unambiguous embedded clauses comprise three-fourths of all embedded clauses; thus they would greatly outnumber ambiguous complement clauses. Moreover, given that the preponderance of embedded clauses in the pilot study provide unambiguous evidence for [T-fin], it is expected that they will be more than capable of counteracting the slight edge that ambiguous complement clauses provide for [T-init].[25]

In summary, the current model does not contain any parameters for extraposition, and thus presents a simplified view of embedded clauses with respect to their position and type (i.e. complement or adjunct). I have focused on how distinguishing complement clauses from other embedded clauses results in a slightly different learning scenario. Nevertheless, the relative infrequency of this type of input and the modest effect it would have if included lead me to conclude that including it would have minimal impact on the overall learning trajectory.

In this section, we have seen how an analytic investigation of parameter interaction gives a principled account of the general trends observed in the child productions. In the initial state, ambiguous input on average pushes the learner to favor T-initial grammars. Further, I have hypothesized that the learner can be pushed all the way to a parameter mis-setting of [T-init]. This corresponds to the analysis of Stage 1, a stage when children consistently produce embedded clauses with an initial T. Once additional parameter setting occurs, the model begins to gradually move toward the target T-final grammar, which corresponds to a gradual decline of non-verb-final embedded clauses in Stage 2 of the children's development. What has not been made precise, though, is how the model expresses the nature of these correspondences. Up to this point the discussion has focused in a general way on how the model learns weights for parameter values. (Indeed, the results in Chapters 4 and 5 for verb-final languages and causatives simply report what weights the model has learned to assign to different parameter values.) There is an intuitive connection between parameter values and the utterances of a language learner,

[25] I have focused here on the role complement clauses in learning TP-headedness, but for reasons analogous to those in the main text, we do not expect adjunct clauses that are right-adjoined to VP (which are also ambiguous with respect to TP-headedness, and which also provide a slight edge to [T-init]) to play much of a role in setting the TP-headedness parameter.

but this connection has not yet been made explicit in the context of the learning model here. In the following section I discuss how I will model child productions given what the model learns for parameter values. This will allow us to make clear predictions about what kind of modeling results characterize the productions of the Swiss German children.

4.4 Predictions for the model

In this section I first introduce a simple way to take results from learning parameter values and using them to model what the actual utterances (including errors) of a learner might look like. In this way we can make predictions for what kind of results to expect when modeling the children. I then present basic predictions for running the model, given the discussion in Sections 4.1 and 4.3.

There is an intuitive connection between the parameter weightings that are learned and the kinds of utterances a learner will produce. Parameter values that are weighted more heavily are more likely to be present in the grammar used by the learner when speaking. We can use this connection to then model which grammar is most likely to be used at any given point in time by the learner when speaking. This can be done in the following simple way.

At any given point in time we can take the probability distributions of the syntactic parameters and sample values from them so as to construct a grammar. This is in fact what we have already seen when the model attempts to learn parameter values. However, instead of comparing this sampled grammar to see whether it is compatible with the input, we can simply assume that this is the grammar that the speaker will use to produce an utterance at that time. In other words, the utterance that the speaker produces at that time (a kind of output) will be compatible with that sampled grammar. To get a more accurate modeling perspective of which grammar the learner will use, we can repeat this sampling process many times and find out on average how likely each grammar is to be used to produce utterances.

We can then compare these averages of the sampled grammars to the actual productions of the Swiss German children. For example, in Stage 1 when the children consistently produce certain kinds of errors, we expect the sampled grammars to be consistent with those errors a high percentage of the time. That is, we expect high averages for these non-target grammars during Stage 1. During Stage 2, we expect the averages of these non-target grammars to decline, while the average of the target grammar increases.

It is important to note that by sampling grammars in this way, we cannot make any predictions about what the speaker will *choose* to say. The speaker might not say anything, or the speaker might not produce any embedded clauses at all. Rather what is being modeled is the following: given that the speaker does produce a particular utterance, that production will conform on average with the grammars that are likely to be sampled.

The question then is what are the precise parametric settings for these non-target grammars. Recall that the core component of the analysis in Sections 4.1 is that children initially favor a T-initial grammar, and that production errors result from V-to-T movement to this head-initial TP. Thus when children make errors, their grammars are T-initial. I further hypothesized in Section 4.3 that parameter interaction not only pushes the children toward a T-initial grammar, but leads to a [T-init] parameter mis-setting. Given this, the prediction is that in modeling the Swiss German children the model will at first consistently sample T-initial grammars (Stage 1) before gradually sampling T-final grammars at an increasingly greater rate that approaches 100% (Stage 2):

(36) *Prediction for TP-headedness*
a. Stage 1: high sampling rate for T-initial
b. Stage 2: sampling rate for T-final gradually increases and approaches ceiling

What about the other parameter settings? Given Schönenberger's claim that the children's productions of matrix clauses are highly target-like, the null hypothesis we can make is that the children are adopting a grammar that is minimally different from the adult grammar. This grammar would be non-target-like in being T-initial, but would otherwise have target parameter settings. If such a grammar in (37) is sampled, it would give rise to error in embedded clauses, but no errors in matrix clauses are expected.

(37) *Hypothesis for sampling non-target grammar (minimally different from target)*
[+VT, +TC, C-init, T-init, V-fin]

(38) *Predicted grammar for convergence (adult target grammar)*
[+VT, +TC, C-init, T-fin, V-fin]

Thus in Stage 1 we expect a high sampling rate for grammar (37). This grammar will gradually be supplanted by the target grammar (38) in Stage 2 as the children approach adult-like performance. Results are reported in the following section.

5. Results and Discussion

In Section 5.1 I first describe the prior probabilities for parameter values and the procedure used to update these priors with posterior probabilities. I then provide results and discussion in Section 5.2. The results confirm the prediction in (36). Initially the model learns a non-target T-initial setting before gradually shifting to a T-final setting. The basic insight of the learning approach adopted here is thus successful in accounting for parameter mis-setting in children's grammars. Accordingly, these results also provide support for a T-final analysis of the adult grammar. Interestingly, the hypothesized non-target grammar in (37) is one of two non-target grammars that are highly sampled by the model. Although somewhat surprising, we can use these results to reach a deeper

understanding of the children's development. This emerges once we take a more detailed look at the acquisition data, which I discuss in Section 5.3.

5.1 Priors and update procedure

The model is run with a program written in the Church programming language (Goodman et al. 2008). I used the *bher* implementation of Church.

For all parameters, I use priors with weak expectations for the learner about the shape of the adult grammar. As discussed in Chapter 2, such priors with weak expectations can be used to model a less tentative learner. Such a learner is less cautious and more is likely to adopt (at least temporarily) a non-target analysis. This is precisely what we want to capture the errors in child Swiss German productions. This is especially helpful in modeling mis-setting TP as being head-initial. Although T-initial is favored initially via parameter interaction, it is only favored with ambiguous input, which is not as impactful as unambiguous input in shifting parameter weights. Further, as the learning process progresses, embedded clauses emerge as more and more likely to be most compatible with a T-final setting. In a sense, T-initial is fighting a losing battle that potentially becomes more lopsided with every subsequent token of input. Still, to model the children's high error rates in Stage 1 of the development, T-initial must have a heavy weighting in the early going. To increase the likelihood of that happening, then, I have used priors with weak expectations.

As the null hypothesis, I assume all parameter values are weighted equally in the initial state. Each parameter will be represented with its own dirichlet distribution, with initial pseudo-count values of .5 for all parameter values. The pseudo-counts are the parameter weights and represent the learner's expectations regarding the shape of the adult grammar. The prior pseudo-count values are thus equal, and are low, which represents a learner with weak initial expectations. Each parameter's dirichlet distribution uses its weights to generate a probability for a given parameter value. These probabilities are then used to sample a grammar for a given token of input. For example, the parameter for VP-headedness will initially have a dirichlet distribution of *dir*(.5, .5), where the first pseudo-count total corresponds to V-initial, and the second to V-final. For more discussion of the dirichlet distribution, I refer the reader to Section 2.4 in Chapter 2.

The update procedure is as follows and is applied equally to all parameters. For each token of input, the model will sample from the parameter weights so as to select 10 (potentially non-distinct) grammars that are compatible with the input. Each of these sampling procedures can be called a chew. These 10 chews represent what the model learns for each token of input. This can be expressed in terms of pseudo-counts. If only one grammar G_1 is selected in all 10 chews, the pseudo-counts of all the parameter values that comprise that grammar will be increased by 1. If G_1 is V-final after the first token of input, then in the example of VP-headedness, the updated pseudo-counts will be the dirichlet distribution of *dir*(.5, 1.5). This means that the adjustment to the weight of any

parameter value being used after a single chew is 0.1. If after the first token of input some input-compatible grammars using V-final were sampled 7 times, while input-compatible V-initial grammars were sampled 3 times, then the updated dirichlet distribution for VP-headedness would be *dir*(.8, 1.2). This process iterates after each subsequent token of input, with pseudo-count values increasing accordingly. However, to expediently see the model shift most of the probability mass onto one parameter value over the other (and thus learn the parameter setting with a heavy weighting) without the pseudo-count values becoming very large, I have chosen to normalize the pseudo-count total of each parameter after a certain amount of input. To model the productions of the Swiss German children, for each parameter I normalized the sum of the pseudo-count values to 5 after every 10 tokens of input. This means that the sum of the two pseudo-count values for, say, VP-headedness would equal 5 after 10 tokens of input. I note that the normalization process can sometimes result in the weights becoming very small. The pseudo-count values for the dirichlet distribution must be greater than 0, but if the value becomes too low, the model will treat it as if it is 0, resulting in a domain error. To avoid this, a parameter's weights were adjusted to 0.02 and 4.98 if after being normalized, they were below 0.02 or greater than 4.98.[26]

5.2 Results

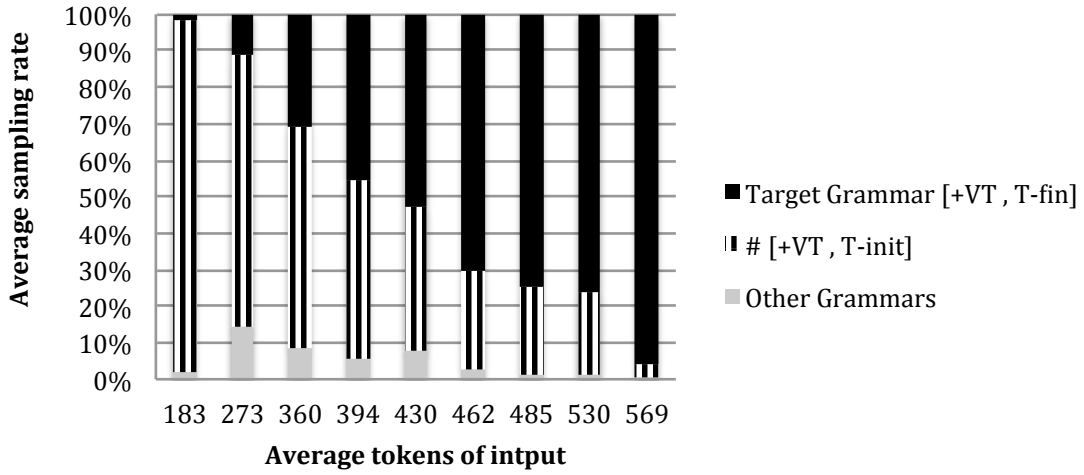
Results illustrating which grammars were sampled with respect to TP-headedness and V-to-T movement are given in Figure 3.2. These results model what type of grammar a learner would use on average to produce an utterance, as per the sampling procedure described in Section 4.4. This sampling procedure was done at various points during the learning process and reflects the shift in weights on different parameter values. During each sampling procedure, the parameter weights were sampled 20,000 times; average rates of which grammars were selected across different runs of the model are given below. These samples thus represent snapshots of the learner's development; taken together they chart the learner's developmental path. The model was run 50 times. Although the model does not struggle to learn the adult grammar, not all runs of the model result in high initial error rates like those in Figure 3.2. Still, by 10 tokens of input, 86% of the runs favor T-initial. This illustrates that the input indeed pushes the learner toward a non-target setting for TP-headedness. Nevertheless, to model the development of the Swiss German children, the model must heavily favor T-initial. The results reported below are averages of 13 runs (26%) of the model with such high initial error

[26] Slightly different adjustments were used when sampling grammars to model utterances, the results of which are given in Figures 3.2 and 3.3. Values of 0.01 and 4.99 were used if the unadjusted normalized values were below 0.01 or above 4.99, otherwise the unadjusted normalized weights were used for this sampling. A lower minimum of 0.01 was chosen because subsequent iterations of the normalization process do not apply to the weights used for this sampling for utterance grammars. It is the normalization process that can bring the weights much closer to 0, and therefore a slightly higher minimum value of 0.02 was used for weights that would subsequently be reinforced and normalized.

rates. These runs, then, can be taken to model the children's development. In Section 7 I discuss the significance of the fact that not all runs have near-ceiling error rates, and how this fact could be an additional strength of the model.

The results in Figure 3.2 match the learning curve of the Swiss German children through Stages 1 and 2 of their development. Early on (around 180 tokens of input) we see a very high sampling rate of [+VT, T-init] grammars. This would result in near-ceiling error rates for verb placement in the relevant embedded clauses. This sampling rate also confirms prediction (36a) and corresponds with Stage 1 of the children's development, during which the children consistently have non-target productions in these utterances. Subsequently, we see the sampling rate of T-initial gradually decline while the sampling rate of the target T-final setting gradually increases; the sampling rate of [+VT] remains high. This confirms prediction (36b) and would now result in variable production of verb-placement errors: when a T-initial grammar is selected, an error is possible, whereas no verb-placement errors are expected under a T-final grammar. By around 570 tokens of input, the model has a very high T-final sampling rate, indicating largely adult-like performance. Again this matches the children's development: Stage 2 of their development is characterized by variable production, with target-like utterances becoming gradually more prevalent until the children are close to adult-like performance at the end of the study.

Figure 3.2 Developmental Course: [27]
Sampling averages of grammars for TP-headedness and V-to-T movement



[27] To compare developmental trajectories across different learning curves, I selected two kinds of points in time to sample from within each run of the model. I first selected 5 points within each run that appeared relatively comparable across runs. Then within each run, I took the midpoint (rounding up to the nearest 10 tokens of input) between each of these 5 points for that run. This gave an additional 4 points for sampling, bringing the total to 9 points for each run. The average values of these points are reported in Figures 3.2 and 3.3.

In sum, a first look at the results indicates a high degree of success for the model. The predictions in (36) are confirmed. That is, the basic insight that the Swiss German children are raising the verb to a head-initial TP is borne out in the results. The model thus accounts for the children's verb-placement errors in embedded clauses, as well as their developmental trajectory as described in Schönenberger's (2001, 2008) study. Further, this success can be taken as support for an analysis of the target grammar as being T-final, as was discussed in Section 4.1.

Above I showed the sampling results for two parameters. Let us now consider the sampling results for entire grammars. These results are shown in Figure 3.3 and Table 3.2. Figure 3.3 and Table 3.2 are simply a more detailed enumeration of the grammars that were sampled and reported in Figure 3.2. By looking more closely at these we see that the hypothesis in (37), repeated below, is not correct.

(37) *Hypothesis for sampling non-target grammar (minimally different from target)*
[+VT, +TC, C-init, T-init, V-fin]

Although the grammar in (37) is highly sampled as a non-target grammar, it is in fact one of the two non-target grammars that are highly sampled by the model. The grammar in (37) is [+TC], and the other non-target grammar that is highly sampled is the same as the one in (37) except that it is [–TC]. On average, only a small proportion of the time is some other non-target grammar sampled, although there is a slight spike around 270 tokens of input. Overall, the results indicate that in the early stages of learning, the model has correctly learned [+VT], [C-init], and [V-fin], and has incorrectly learned [T-init], but has not yet settled on a setting for [±TC]. Initially, the non-target [–TC] grammar has a higher sampling average than the non-target [+TC] grammar, but as the overall non-target grammar sampling rate goes down, it is the non-target [+TC] grammar that is the more sampled non-target grammar.

Figure 3.3 Developmental Course:
A closer look at non-target grammars sampled

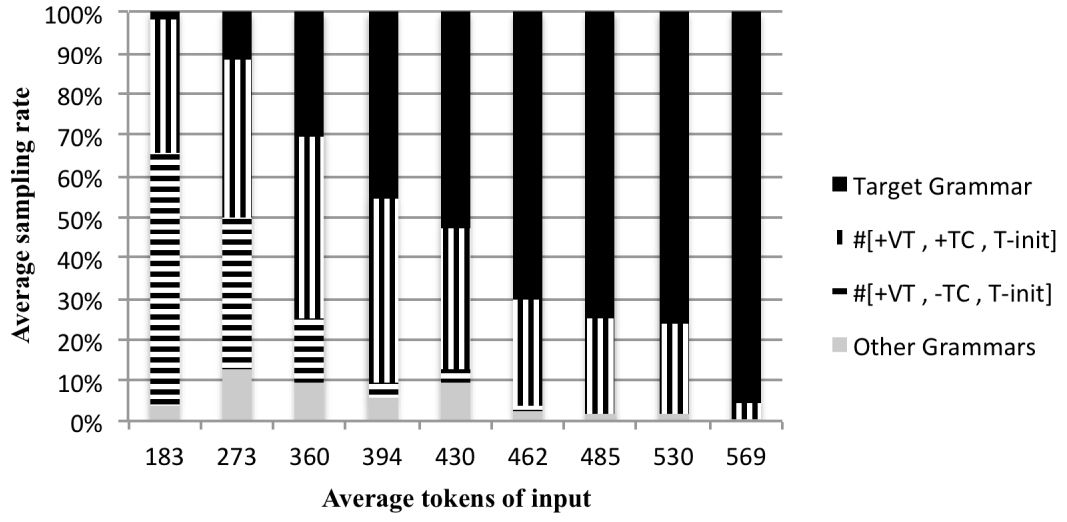


Table 3.2 Developmental Course:
Sampling averages of grammars used to produce an utterance

Average tokens of input	#[+VT, –TC, T-init][a]	#[+VT, +TC, T-init][a]	Total # (Columns 2+3)	Target Grammar	Other Grammars
183	.6188	.3236	.9424	.0181	.0395
273	.3726	.3845	.7571	.1145	.1284
360	.1584	.4513	.5979	.3073	.0948
394	.0388	.4487	.4876	.4530	.0594
430	.0355	.3428	.3783	.5257	.0960
462	.0097	.2603	.2700	.7034	.0266
485	.0024	.2339	.2363	.7477	.0160
530	0	.2225	.2225	.7606	.0169
569	0	.0384	.0384	.9554	.0062

a: Parameter values not indicated in columns two and three are target values (i.e. C-init and V-fin).

Sampling a [–TC] grammar plays an important role early on in having high error rates, but does sampling a [–TC] grammar mean that the probabilistic learner is not accurately modeling the development of the Swiss German children? A high sampling rate for a [–TC] grammar comes as surprise given Schönenberger's claim that the children's productions of matrix clauses are highly target-like. If the model has not yet learned to raise the verb to C, we might expect additional production errors. The following discussion addresses this issue from two angles. First, a [–TC] grammar can still account for all the children's productions in both embedded and matrix clauses. Second, there is some evidence suggesting that the children might indeed be using a [–TC] grammar initially. This evidence comes from two sources: some matrix clause errors that are attested infrequently, as well as a more careful examination of the distribution of different types of subjects in embedded clauses, which I discuss in Section 5.3. In this way we can use the modeling results to arrive at a deeper understanding of some puzzling acquisition data that Schönenberger reports.

An important first observation is that with a [–TC] grammar that is also [+VT, T-initial] we do still account for the attested production errors in embedded clauses. The core of the analysis of these errors is simply that the finite verb is raising to an initial T. Nothing in this analysis hinges on raising the verb to C. Thus the results in Table 3.2 also match the learning curve of the children in Schönenberger's study. Again, early on we have near-ceiling error rates for raising the verb to an initial T (Stage 1 of the children's development). As the learning process continues, more and more weight is gradually shifted to T-final, representing a period of variability between target and non-target productions (Stage 2). By the end of the sampling period, the model has largely learned the adult grammar.

Second, a [–TC] grammar that is also [+VT, T-initial] is compatible with all the target-like matrix clauses that Schönenberger samples from Moira's corpus. Recall from Section 4.3 that all matrix clauses are compatible with this grammar. The verb will raise to T, and the initial constituent can appear in SpecCP (or possibly also SpecTP for

subjects). Further, I have been assuming (cf. Section 4.1) that all post-verbal constituents in the middle-field can appear adjoined to VP (Rosengren 2002). This non-target grammar, in which the verb raises to T, can thus accommodate clauses produced by the children that contain multiple constituents in the middle-field:

(39) Mengisch törf t'Eliza ned i mis Zimmer inecho und
sometimes may the-Eliza not in my room in.come and
mengisch scho. (M: 3;10)
sometimes yes
'Sometimes Eliza isn't allowed to enter my room, and sometimes she is.'
(Schönenberger: 212)

The question that adopting a [–TC] grammar raises, though, is whether children will overgenerate non-target matrix clauses. The concern is that because the verb is not raising to C, we will lose the V2 nature of Moira's matrix clauses; we might expect the verb to appear in third (or even fourth) position. This could occur with constituents co-occuring in SpecCP and SpecTP (or adjoined to TP):

(40) #[$_{CP}$ XP [$_{TP}$ (YP) [$_{TP}$ ZP [$_{TP}$ V+T [… ~~V~~]]]]]

However, there are several points that can be made for the case of the children using a [–TC] grammar. To begin, I repeat an important point from the previous Section 4.4 regarding predicted utterances in this model. No specific utterances are predicted to be produced *per se*: the model does not predict any productions with the verb in third position. Utterances such as (40) are not entailed by a [–TC] grammar; such an utterance depends on the speaker making additional choices, such as filling both SpecCP and SpecTP. Rather, we expect utterances such as (40) with a frequency that is determined by the joint occurrence of all these choices. It is possible that independent factors considerably reduce the likelihood of making all these choices, resulting in a situation where utterances such as (40) are rare. This is the tack I will take when I explore this issue in more detail in Section 5.3. The most conservative position then is to say that the children's adopting [–TC] grammar is plausible because such a grammar can account for all the utterances they do produce. By way of contrast, it is less plausible to claim that the children have adopted a C-final or V-final grammar: not all child productions are compatible with these non-target parameter settings.

Moreover, there is some evidence showing (a) utterances such as (40) are produced, albeit rarely; and (b) some independent factor could be playing a role in the rarity of these utterances. As regards the first point, in Section 2.2 I mentioned that Schönenberger observes a handful of production errors in the sample of Moira's matrix clauses in which the finite verb does appear in third position. An example is given in (41).

(41) %Nämlich ned alli Lüt händ di gliiche Schtimm. (M: 4;11)
particle not all people have the same voice
'Not everybody has the same voice.' (Schönenberger: 198)

An example such as (41) is precisely the kind of production that would be allowed by a [–TC, +VT] grammar, but is unexpected in a [+TC, +VT] grammar in which the finite verb would be preceded by only what is in SpecCP.

The rarity of these examples is in line with results from Waldmann (2011; see also references therein) who investigates the frequency with which Swedish children have non-target productions of the finite verb in third-position. Waldmann finds that these errors are attested in spontaneous speech across children, although for some children they are consistently reported to be rare. Further, for one child that Waldmann reports on in detail, errors of this sort are relatively common before age 3;06, but become noticeably rarer after this age.

Turning to standard German, Gawlitzek-Maiwald et al. (1992) report several examples of spontaneous production errors in which the finite verb appears in third position in matrix clauses. These errors appear around age 3;0.

In light of these observations, we can note that Moira's corpus begins only at age 3;10. The sample size of Moira's matrix clauses was less than one-fourth that of her embedded clauses, and it is possible that a larger sample size starting from an earlier age would reveal more errors of this type.

There is also some evidence suggesting that productions like (40) might be rare due to some independent factor, which I discuss in the following section. This will involve a more careful examination of the acquisition data that reveals a subtler developmental trend that is consistent with the children having adopted a [–TC] grammar.

In sum, the results here show that the model is largely successful in modeling the acquisition of verb placement in the Swiss German children. This in turn provides support for an analysis of the adult grammar as being T-final. The non-target grammars that are sampled are T-initial as predicted in (36). Primarily only one other non-target grammar is sampled in addition to the hypothesized non-target grammar in (37). These non-target grammars are T-initial and may be –TC, but otherwise have parameters with target values. The rate at which these non-target grammars are sampled also matches the developmental stages of the children. Further, other non-target grammars are sampled infrequently. Finally, after additional learning the model closely approaches the target adult grammar, as per (38).

5.3 A closer look at the acquisition data:
The distribution of subjects in embedded clauses

In Section 5.2 I discussed various reasons why a [–TC] grammar is compatible with the development of the Swiss German children. One possible objection was the issue of the children over-generating non-target productions of the form in (40), repeated here, in which the finite verb occurs in third-position in the matrix clause. As discussed, such non-target productions are in fact attested, albeit rarely.

(40) #[$_{CP}$ XP [$_{TP}$ (YP) [$_{TP}$ ZP [$_{TP}$ V+T [… ~~V~~]]]]]

There is also some evidence from embedded clauses suggesting that the Swiss German children have a tendency to avoid a syntactic configuration that would result in a matrix clause of the sort in (40). What (40) shows is that to have the finite verb in third position, there must be a constituent in SpecCP and some other constituent in SpecTP or adjoined to TP. A more careful examination of some puzzling acquisition data involving embedded clauses can lead us to conclude that, at least in earlier developmental stages, the children avoid having any constituent in an embedded SpecTP or adjoined to an embedded TP. This is a surprising fact, but it is an observation that is independent of whether the grammar is [–TC]. If this tendency exists in embedded clauses, we can suppose that the same holds for matrix clauses. The conclusion is that children initially avoid TP for any constituent other than what is in T^0 (i.e. they avoid SpecTP or adjoining to TP). Whatever the cause of this might be, we can take this conclusion as supporting, or at least as being compatible with children's initially adopting a [–TC] grammar. And while this phenomenon still remains unexplained, this conclusion gives us a precise characterization of the data that brings it in line with the broader results of the learning model.

Let us now take a closer look at the acquisition data. The core production patterns are repeated below.

(12) *Verb placement in embedded clauses in Stage 1 of child productions*
a. #… Complementizer Subject V_{fin} X …
b. #… Complementizer V_{fin} {Subject X / X Subject} …
c. #… Complementizer V_{fin} Subject
d. … Complementizer Subject V_{fin}

According to the analysis here, the finite verb raises to T in the children's embedded clauses. Assuming that the complementizer is in C, then it is clear that no other constituent is in SpecTP in (12b, c). In (12a, d), though, there is an intervening constituent – the subject. Nevertheless, it is still possible to take (12a, d) as being compatible with the conclusion above concerning additional material in SpecTP, when we take a closer look at these subjects. Interestingly, as Schönenberger (2001) discusses,

up until age 4;11 this intervening subject is virtually always a stressless subject pronoun. In the adult grammar, this is the only position (i.e. the Wackernagel position immediately after the complementizer) that these pronouns can appear in in embedded clauses. Before age 4;11, the subject after the finite verb (12b, c) can be either a stressless subject pronoun or a fully phrasal non-pronominal DP. It is only after age 4;11 that non-pronominal subjects begin to precede the finite verb with any regularity.

These word order patterns before age 4;11 are illustrated in the tables below, which are based on Figures 6a and 6b in Schönenberger (2001: 278). Tables 3.3 and 3.4 are based on Moira's corpus of embedded clauses introduced by the complementizers *wenn* 'when, if', *dass* 'that', and *öb* 'whether, if'. They show the distribution of non-verb-final embedded clause productions between ages 3;11-4;11 in which (a) a pronominal subject ($Pron_{Subj}$) immediately precedes or follows the raised verb; (b) a non-pronominal subject (DP_{Subj}) immediately precedes or follows the raised verb; (c) some other constituent X immediately precedes or follows the raised verb. The occurrence of an immediately post-verbal non-pronominal subject is fairly evenly distributed month to month in this period, but it is almost never attested pre-verbally. During this period of development, the only constituent that regularly appears between the complementizer and the verb is a pronominal subject. After age 4;11, non-pronominal subjects begin to occur more frequently before a non-final verb; Schönenberger (2001: 280) reports 16 examples of this with a non-pronominal subject and 37 additional examples with some other constituent X between the complementizer and a non-final embedded verb.

Table 3.3 Constituents immediately following the finite verb in non-verb-final embedded clauses introduced by wenn, dass, and öb in Moira's corpus, ages 3;11-4;11

Type of production by post-verbal constituent	**Number of productions**
[…Comp. V_{fin} $Pron_{Subj}$ X…]	259
[…Comp. V_{fin} DP_{Subj} X…]	84
[…Comp. V_{fin} X…]	22

Table 3.4 Pre-verbal constituents in non-verb-final embedded clauses introduced by wenn, dass, and öb in Moira's corpus, ages 3;11-4;11

Type of production by pre-verbal constituent	**Number of productions**
[…Comp. $Pron_{Subj}$ V_{fin} X…]	100
[…Comp. DP_{Subj} V_{fin} X…]	2
[…Comp. X V_{fin} …]	1

Further, productions of the sort in (15), repeated below, where a pronominal subject co-occurs with some other constituent between the complementizer and a non-final finite verb, appear to be infrequently attested before 4;11 (Schönenberger 2001: 146-147; 306-

307).

(15) # [wenn du no einisch wärsch do drin] (M: 4;11)
if you still once would-be there in
'if you were still once in there' (Schönenberger: 307)

What emerges is that in both the adult and child grammars, stressless subject pronouns have a special distribution that distinguishes them from other subjects. In the adult grammar, no constituent can intervene between a stressless subject pronoun and the embedded clause's complementizer. Before age 4;11 in the children's grammars, a stressless subject pronoun is nearly always the only consistuent that can intervene between the raised verb and the embedded clause's complementizer. The parallel between these two phenomena is striking, and it is plausible that the privileged position of stressless subject pronouns in both cases is related. This is the approach I will pursue below.

We can make sense of the different distribution between stressless subject pronouns and other subjects by appealing to the difference between syntactic heads and phrases. Moreover, this difference will allow us to maintain the claim that initially nothing is in SpecTP or adjoins TP in the child grammar. I assume that stressless subject pronouns are heads that must head-adjoin to C in the adult grammar (cf. Zwart 1997 and Haider 2010). In matrix clauses, this means they immediately precede or follow the finite verb (assumed to be in C), depending on whether the clause is subject-initial or not. In embedded clauses, these subjects immediately follow the complementizer. As a result of head-adjunction, then, phrasal material cannot intervene between the pronominal subject and either a finite verb or complementizer in C in the adult grammar. If the Swiss German children have learned that these pronominal subjects require a host to adjoin to as in the adult grammar, then the pronominal subjects in (12a, d) are not in SpecTP; rather they are in C with the complementizer.[28] In the framework of the analysis here, non-pronominal subjects, as full phrases, could appear between the finite verb and the complementizer only by being in SpecTP (or adjoined to TP). Given that non-pronominal subjects rarely intervene between the two before age 4;11, it follows that until age 4;11 the children nearly always leave SpecTP empty in embedded clauses. Beginning from age 4;11, when non-pronominal subjects (or some other constituent) begin to precede the raised verb somewhat more frequently, I will assume that children have begun to merge these non-pronominal constituents in SpecTP (or otherwise adjoining them to TP) more frequently. Thus we can account for the privileged position of these pronominal subjects in the

[28] There remains an additional acquisition puzzle concerning productions involving stressless subject pronouns that follow the raised finite verb in embedded clauses. These are non-target; in the adult grammar these pronouns must immediately follow the complementizer. This is an interesting puzzle that goes beyond the scope of the discussion here. One possible direction to pursue is that when the children raise the verb in embedded clauses they have transferred something they have learned from matrix clauses: these same pronouns immediately follow the raised verb in matrix clauses when the subject is post-verbal.

children's productions by hypothesizing that the children have adult-like pronominal adjunction to C. Further, with this hypothesis we can maintain the conclusion that only material in T is in TP in the children's earlier productions. Finally, I note that this revised analysis of subjects does not substantively alter the basic analysis of the children's non-target productions in (12a-c): whenever there is an error with the embedded finite verb preceding some other constituent, this results from raising the verb to a head-initial T.

Such a conclusion about material in TP in the children's embedded clauses also points to the beginning of an account of why utterances of the form in (40) are rare. If we assume that there is a strong tendency to avoid material in TP other than what is head-adjoined to T, then we expect matrix clauses with TP material and a third-position verb to be rare as well. At age 4;11 this tendency begins to fade, and there is an increased occurrence of non-verbal constituents adjoined to TP in embedded clauses. If the increased frequency in adjunction to TP holds across the board, in embedded and matrix clauses, we do not necessarily expect to see a con-current spike in the frequency of matrix clauses like (40), with the finite verb in third position. Such a spike is not expected if the use of a [–TC] grammar is sufficiently on the decline after age 4;11.

Seen from this perspective, there is support for such a correlation between the increase in adjunction to TP and the decline of a [–TC] grammar. After age 4;11 is also when verb placement errors in child productions in embedded clauses begin to more rapidly decrease (cf. Table 3.1), indicating that Stage 2 is in full swing. My analysis of these target embedded clause productions has been that the clause-final verb is in a head-final TP. The finite verb in the matrix clauses of these productions still occurs in second position. As TP is head-final in these utterances, we can conclude that the verb in matrix clauses is in C. This means that the increase in target embedded clauses around age 4;11 correlates with an increase in the use of a [+TC] grammar. And the stronger the increase in target embedded clauses, the stronger the increase in [+TC]. Conversely, the onset of much greater target-like performance is a time when the use of all non-target grammars decreases, including a [–TC] grammar. We can thus correlate this time in development with both a decrease in the use of a [–TC] grammar and with an increase in material adjoined to TP. A further expectation is that the non-target productions during Stage 2 predominantly result from sampling the [+TC] grammar in (37). This is expected if the correlation holds between a decline in the use of [–TC] grammar and the appearance of additional phrasal material in TP. Indeed the results of the model match the general contours of these correlations. In Table 3.2 we saw that although the non-target [–TC] grammar has the highest average initially, as the productions begin to become more variable in Stage 2 of development, the sampling rate of the non-target [–TC] grammar drops, and the [+TC] grammar becomes the most sampled non-target grammar. In sum, the discussion here reveals a broader developmental trend that goes beyond a puzzling fact about the distribution of embedded subjects and allows us to incorporate the data into the more general results of the model.

If this interpretation of the acquisition data is on the right track, it still remains an open question as to why children would have such a tendency early on regarding TP. Nevertheless, I will take this as support that children could be using a [–TC] grammar while producing embedded clauses with verb placement errors. Moreover, the results from the model allow us to take an observation about the distribution of embedded subjects and posit a more far-reaching developmental trend that cuts across matrix and embedded clauses.

6. Comparison with other learning models

In this section I briefly compare how the other learning models introduced in Chapter 2 fare with the acquisition puzzle in Swiss German. The general discussion in Chapter 2 provides a more thorough comparison, and many of the comparative remarks in Chapters 3 and 4 hold here as well. Here I will limit myself to brief summary comments in light of the specifics of Swiss German as it has been presented.

The Swiss German acquisition data pose a challenge for models such as Sakas and Fodor (2001) that learn only from unambiguous input. This kind of a model is designed to converge on the target grammar via triggering input, but the challenge for this model is accounting for the learner using non-target grammars. One possibility that this model can avail itself of is to use default parameter settings. We could suppose that in the initial state the learner has a T-initial parameter setting. Learning additional parameter settings and receiving input of embedded clauses would ultimately cause the learner to adopt a T-final grammar. Nevertheless, the variability we see in Swiss German child productions remains a challenge. If the learner is producing target utterances, then according to this model they must have received unambiguous input as to what the target grammar is. But in that case, why would the learner ever revert to a non-target grammar by producing a non-target utterance? To account for the within-speaker variability during acquisition, a further modification would be to say that there is a probabilistic component (cf. Sakas and Fodor 2012: 91). So modified, Sakas and Fodor's Structural Triggers Learner (STL) would learn only from unambiguous evidence, and would have a T-initial default, but would incrementally replace this default value with T-final. More specifically, the default would amount to a very strong probabilistic weight in the initial state for T-initial. If the learner sampled a grammar to produce an utterance in the initial state, we would see error rates near 100%.[29] After setting [+VT] via unambiguous evidence, the STL parser would detect unambiguous evidence for T-final in some embedded clauses. Upon encountering these embedded clauses, the learner would incrementally increase the weight for T-final until the parameter has been set. In the interim, while the weights are shifting from being

[29] Thus even with a probabilistic component, the STL cannot initially be 'agnostic' with 50-50 weights on competing parameter values. Such a learner would progressively move toward the target state, and under the analysis of Swiss German here, would never sample the non-target grammar with near-ceiling frequency as has been claimed the Swiss German children do.

concentrated on T-initial to T-final, the learner would be able to sample either value, which results in variable non-target productions. Such a proposal is quite a departure from the original formulation, but it still would learn only from unambiguous evidence, and would seemingly account for children's variable error rates. Nevertheless, this modified STL still predicts that we only see child errors that reflect the default value. Because of its reliance on unambiguous evidence, the STL only moves toward target parameter values; non-target values, and thus errors, must result from the default. As was mentioned in the Chapter 1, though, if we take a cross-linguistic perspective, we see error patterns that cannot all be accounted for with a single default. I discuss this in more detail in Chapter 6, and suggest that these errors can be understood with a probabilistic model that learns from ambiguous evidence.

As we saw in Chapter 2, the Triggering Learning Algorithm in Gibson and Wexler (1994) is able to model variable productions in an input-driven way. Given the right sort of ambiguous input, the model can be pushed back and forth between multiple non-target grammars. Additional input could then allow the learner to adopt the target grammar, but once the learner has adopted the target grammar, the learner will not deviate from that. I will not speculate on what kind of input would allow the Swiss German children to move back and forth between multiple non-target hypotheses. Supposing such input did exist, this scenario with the TLA still makes a prediction that does not receive any clear support. The prediction is that throughout all of Stage 2 of the children's development, the stage during which we see variable productions, all variability is the result of the children exclusively using multiple non-target grammars. The results in Section 5 do show that, at least in the initial part of Stage 2, the children are considering multiple non-target grammars. But the results show that the remainder of Stage 2 is composed of competition between the target and a single non-target grammar. It is of course possible that when the children are frequently producing target embedded clauses they are adopting some grammar that is non-target-like in some other respect. However, there is no evidence presented in Schönenberger (2001, 2008) of other well-attested errors from throughout Stage 2 that are independent of verb placement in embedded clauses.[30] Lacking such evidence, the TLA is faced with the challenge of determining what the other non-target grammar is during Stage 2 and providing evidence for it. The model in Gibson and Wexler thus shows a similarity to that in Sakas and Fodor: adopting the target grammar brooks no further competition with it. The probabilistic model I have proposed differs in that adopting the target grammar at any given moment is only part of the learning process that gradually shifts more and more weight to the target hypothesis.

This gradual learning of the target grammar is precisely what we see in Yang's

[30] As mentioned in notes 1 and 2, Schönenberger (2008: 103-106) suggests that later in development, Moira might have overgeneralized the application of VR/VPR in a small subset of non-target embedded clauses. If correct, this might provide the TLA with another mis-set parameter. However, I have already questioned in note 1 whether Schönenberger's suggestion is correct, as it is not clear that these errors result from anything more than a non-target [T-init] value.

(2002) probabilistic learning model, and in many core respects Yang's model is similar to what has been presented in this chapter. The learning we see here via parameter interaction is fully expected under Yang's model but goes beyond any of the cases discussed by Yang, who essentially focuses on parameters in insolation. This learning via parameter interaction is thus a previously unexplored way to learn. The work here serves to draw attention to how parameter interaction plays a pivotal role in the learning process. In the following chapter we will see that this parameter interaction helps the learner converge on a target-consistent grammar. In this chapter, parameter interaction instead contributed to mis-learning by pushing the learner initially in the direction of a non-target parameter setting. In both cases the evidence that played a crucial role was ambiguous input, which is also an important source of evidence in Yang's model.

7. The broader German perspective

In this section I consider a primary issue for further research regarding the learning model and other varieties of German. This relates to the results of the model and acquisition data from other German dialects with respect verb placement in embedded clauses.

A question that arises when we go beyond the Lucernese Swiss German acquisition data is how well the learning model generalizes to the reported acquisition findings in other varieties of German. This question is especially relevant given the early claims in the literature that children learning standard German do not make errors with verb placement in embedded clauses (Clahsen and Smolka 1986; Mills 1985). This is admittedly a coarse comparison because the dialects of German have a range of fine-grained differences, and these differences have also not been worked into the learning model. Still the dialects all share the same core properties of finite verb placement: V2 in matrix clauses and verb-final in embedded clauses. Based on the studies cited in Schönenberger (2001), it is possible to identify some emerging trends in the growing body of acquisition literature that point to a variety of acquisition trajectories across children. However, given the limited nature of the data (both the number of children studied and the size of the children's corpora of spontaneous productions) any conclusions must currently remain speculative. Nevertheless, a possible strength of the learning model in this chapter is that it does not predict all children to pattern like the Lucernese children.

The studies cited by Schönenberger fall into two categories. Some report either no errors or only a handful of errors in embedded clause verb placement in spontaneous productions. Other studies suggest there are further cases of prevalent learner errors in addition to Lucernese, which has been the focus of this chapter. I review these two groups of studies below.

Several studies of children learning standard German (Clahsen 1982; Müller 1993; Rothweiler 1993) that are cited by Schönenberger report either no errors or a very small proportion of errors in embedded clause verb placement. To these can be added similar

findings from the study in Penner (1990) for Bernese Swiss German and from Schönenberger's own analysis of a small corpus for Zurich Swiss German. The errors that are reported are of the familiar non-verb-final variety that we have seen for Lucernese. The conclusion that error-free embedded clauses is a widespread phenomenon in child German seems somewhat premature given the small sample size of the corpora under consideration. With the possible exception of the longitudinal study in Penner (1990), the corpus for each child in these studies is no more than several hundred for each child (Rothweiler 1993) and sometimes as small as 12 (Clahsen 1982). By way of contrast, Moira's corpus of embedded clauses for Lucernese numbers around 5000. It is thus possible that larger corpora would reveal higher error rates.

In contrast, other studies cited by Schönenberger report a much higher frequency of errors in embedded clause verb placement. One child learning standard German in Gawlitzek-Maiwald, Tracy, and Fritzenschaft (1992) has a near-ceiling error rate of non-verb-finality in embedded wh-clauses, although again the sample size is small. The longitudinal study in Penner (1996) is reported to have a large number of embedded clause productions from a single child learning Bernese Swiss German (a majority of its corpus of 1100 utterances). During this child's development, there is a stage during which the child varies between final and non-verb-final embedded clauses, before gradually producing more and more verb-final embedded clauses. This resembles Stage 2 that we saw for the children learning Lucernese Swiss German.

In sum, it is difficult to draw any sure conclusions about embedded clause verb placement in German more generally, but it is possible to speculate about three emerging generalizations from the literature. I speculate that (i) some children have few or no errors; (ii) for some children, they are frequent (cf. Penner 1996); and (iii) for some children error rates initially approach a ceiling before declining, as we saw with Lucernese. It is even possible that children representing all three developmental paths can be found in all German dialects given a large enough number of children studied.

That three distinct developmental paths can be found among German children is compatible with the results reported in Section 5. The results reported in Section 5 model the third developmental path above, one with very high error rates, which is found with the Lucernese children in Schönenberger's study. It is important to repeat that these results are averages only of runs of the model that were consistent with such a developmental course. This occurs 26% of the time. Although the model favors T-initial in the early going 86% of the time, in some runs of the model the learner receives only a moderate or weak initial push toward T-initiality (cf. discussion in Section 4). With only a weak initial push, say, toward T-initiality, the learner can settle on a heavier weighting for a T-final parameter setting at an earlier point in the development course. Thus it is entirely possible that by the time the learner begins producing embedded clauses, a T-final setting has already been acquired. Consequently, a range of different frequencies below the high rates reported in Section 5 are possible for sampling a T-initial grammar

when modeling speaker productions. This range is based on how strong the initial push toward T-initial is: the stronger the push, the higher the expected error rates. A possible strength of the model, then, is its flexibility to model varying developmental paths across children: not all children have initial high error rates, and not all runs of the model have such high error rates. The empirical range of the model could thus be further tested with additional in-depth acquisition studies of children's embedded clauses.

8. The relation between input and learning

Having looked in detail at how input plays a role in how the model learns verb placement in Swiss German, I now discuss more general considerations regarding the relation between the input and how the model learns. I will focus on two issues here. The first concerns the frequency of certain kinds of evidence and learning. A core component of the learning model is that input frequency plays a direct role in learning outcomes and trajectories. Second, I discuss the issue of uptake, or how much input the model actually learns from at any point in the learning process. In this work I have simply assumed that the uptake is a single sentence in its entirety. However, nothing about the architecture of the model is committed to this view of uptake, and I discuss several other possibilities and what predictions they lead to.

I begin with the issue of input frequency. We have seen that the model is conditioned in response to the input: different types of input result in different sets of parameter values being reinforced. The more frequent those types of input are, the more likely the model is to learn the parameter values that are conditioned by that input. A direct consequence of this is that input frequency plays a central role in a learner's development.

Various research on diachronic change illustrates that input frequency is related to language learning, and until shown otherwise, it is reasonable to assume that a model for language acquisition would also be sensitive to input frequencies. For example, in the middle of diachronic change, forms that are increasing in frequency in earlier generations of learners are attested more frequently in subsequent generations of learners, whereas forms that are decreasing in frequency in earlier generations are attested less frequently in subsequent generations (Kroch 1989). An example of this sensitivity to input frequency regarding a word order parameter change in the history of Old English is modeled with a probabilistic learner in Pearl (2007). Thus it is plausible to have parametric differences attributed to statistical frequencies of different types of input.

Relating parameter setting to input frequency includes not only cases where a learner's end-state reflects certain input frequencies (as in the case of diachronic change), but also cases of parameter mis-setting and resetting during the course of acquisition. I discussed this in the context of developmental trajectory in Section 4, where the frequency of embedded clauses was correlated with parameter setting. Thus the lower the overall frequency of embedded clauses in the input corpus is, the more likely the model is

to mis-set the parameter for TP-headedness, and the higher the overall frequency is, the less likely the model is to mis-set this parameter. However, it is not just overall frequency that plays a role in development, but also the order in which the learner encounters different types of input. Two learners might encounter the same overall number of embedded clauses, but one of them might encounter a larger proportion of embedded clauses earlier in the learning process. The prediction is that earlier exposure to a larger number of embedded clauses decreases the likelihood of parameter mis-setting, whereas lack of such exposure increases that likelihood. Similarly, in a scenario where there has been parameter mis-setting, then a subsequent exposure to a large proportion of embedded clauses during a particular stage of learning can result in parameter resetting. Preliminary results from the pilot study mentioned in Section 4.3 point toward these predictions being borne out. The pilot study kept track of what types of input were being randomly presented to the learner, and indeed higher error rates were correlated with relatively fewer embedded clauses earlier in the learning process. A natural place to further test this prediction is in detailed longitudinal studies of child directed speech in (various varieties of) German. To my knowledge this has not been done, but the prediction is higher error rates in child productions correlates with lower frequencies of embedded clauses in the input.[31]

I next look at the issue of uptake, or how much of the input the model actually learns from. We can think of uptake as being related to a filter on the input used for parameter setting. I have assumed that there is no such filter on the input. The model learns from every sentence of input, and also from the entirety of every sentence of input, including embedded clauses. Thus in the simulations of the model here, every time the learner encounters a sentence of input, that sentence is the learner's uptake for a particular iteration of the learning procedure (i.e. for a single iteration of sampling for the prescribed number of chews and subsequent reinforcement of choices). As I discussed in Section 4, this contrasts with Lightfoot's (1989, 1991) Degree-0 hypothesis, which acts as a filter on the input. According to this hypothesis, the learner's uptake is limited to matrix clauses and just the edge of embedded clauses, which constitute the domain of Degree-0 learning. To my knowledge, there is no empirical evidence that learners, children for example, always filter certain sub-structures of embedded clauses from their uptake.[32] Thus we can view learning from input that is unfiltered (i.e. the uptake equals the input) as the null hypothesis for a learning model. This is the approach I have taken with the learning model here, and we have seen that the model has success at modeling the learning trajectory of the Swiss German children.

Nevertheless, it is certainly plausible that a learner's uptake may vary over the course of acquisition. For example, earlier in development a learner's uptake may be

[31] In the context of the research program here, a topic for future research is to see whether an increase in the frequency of ambiguous evidence can be used to model a diachronic change with the learning model proposed here.

[32] Artificial language learning experiments could be used to investigate this question.

more likely to exclude (parts of) embedded clauses, perhaps for developmental reasons. A strong version of this, but still weaker than Lightfoot's claim, is that early in development a learner's uptake is restricted to the Degree-0 domain, and that later in development the learner's uptake gradually expands so as to include the entirety of embedded clauses. An important point to observe is that this modified version of the Degree-0 hypothesis, and indeed any kind of filtering of the learner's input, is compatible with the model's learning procedure. When I introduced the learning procedure in Chapter 2, I did so in a context in which each sentence of input constitutes the learner's uptake for a particular application of the learning procedure. However, the learning algorithm can be applied to any kind of uptake, so long as that uptake is well defined. If the uptake is only a sub-part of sentence, then the generative version of the model only needs to generate a string-meaning pair that matches that sub-part. Similarly, the discriminative implementation would simply check the compatibility of a set of parameter values with that sub-part.

Given that the model could operate with a more restricted view toward a learner's uptake, how might we expect the model to fare with more limited uptake in the case of Swiss German, in which embedded clauses in the uptake played a key role? First, I repeat the observation from Section 4 that given the set of parameters and the analysis of Swiss German in Section 4, the structure of embedded clauses beyond the Degree-0 domain must at some point be part of the learner's uptake in order to learn the adult grammar. Only by looking at structure below T in embedded clauses would the model have evidence for learning a target setting of T-final. For this reason, I rejected Lightfoot's hypothesis that the learner's uptake is always restricted to the Degree-0 domain. Still, we can consider a less restrictive hypothesis, along the lines of the discussion above. To take an example, suppose the learner's uptake is initially restricted to the Degree-0 domain, and that later in development this restriction no longer holds. Under this hypothesis we can still model the learning trajectory of the Swiss German children. Recall that nearly all types of matrix clauses favor T-initial via parameter interaction. This is still true if we include in the uptake the edge of embedded clauses from SpecCP to T. Thus so long as the uptake is restricted to the Degree-0 domain, we expect the learner to be pushed toward a non-target T-initial grammar. If the uptake is restricted for a sufficiently long enough period of time, then we expect some runs of the model to have [T-init] parameter mis-setting, as per the discussion in Section 4. Further, the longer the uptake is so restricted, the longer it will take the learner to reset the parameter to T-final. However, once the restriction no longer holds, then given a stretch of input with a sufficient amount of embedded clauses, which favor T-final, the learner will be able to recover and reset the parameter to T-final. Varying the duration of the restriction on the uptake could even be a way of more accurately modeling the time-course of acquisition. In the proof-of-concept illustration in this chapter, I have not attempted to simulate the children's acquisition path in anything approximating real-time acquisition in months and years. As a speculative

remark on how one might begin to go about doing so, we can suppose that if learners persist in maintaining a non-target T-initial grammar for a lengthy period of time, this could be linked to how long the restriction on the uptake (to the Degree-0 domain) holds. In sum, some restriction on the uptake is consistent with the approach taken in this chapter and the results I have presented. The extent to which language learners do in fact have such a restriction remains a topic for careful experimental work.

Related to the question of uptake is the issue of batch learning. We can think of a batch as how much uptake is stored before the learning procedure is applied to it. In the implementation of the model I have presented, an iteration of the learning procedure applies after every sentence of input on a token by token basis. We can say that this implementation has a batch size of 1: the learner stores a single sentence before learning from it, after which the learner no longer stores that sentence. Having a batch size of 1 is a fairly minimal and conservative (both computationally and psychologically) assumption. And if the uptake is minimally a sentence in size (i.e. the uptake is not a sub-part of a sentence) a batch size of 1 is also the null hypothesis. Again, it is certainly plausible that the learner at least sometimes has a batch size that is greater than 1, and having larger batch sizes is entirely compatible with the learning model. For example, suppose that at some point during acquisition the learner has a batch size of 2. This means that the model will store 2 consecutive sentences of input. The model keeps track of the size of this store. In the generative implementation, the model will then sample a single set of parameter values in an attempt to use this same set of parameter values to output 2 sentences that match the sentences of the batch. In addition to this single sampling of parameter values, the learner will make 2 different sets of choices (one per sentence) concerning phrase structure rules so as to generate 2 sentences of output. The first sentence of output is conditioned to match the first sentence of the batch, and the second sentence of output is conditioned to match the second sentence of the batch. If either of these sentences does not match, then the learner will make another attempt by sampling another set of parameter values until the stored corpus of the batch is matched by the output.[33] Similarly, in the discriminative implementation of the model, the learner will attempt to sample a single set of parameter values that is compatible with both sentences in the batch. What learning from larger batches does, then, is to force the learner to use the same set of parameter values to account for a larger chunk of input.

What effect on learning does a larger batch size have? Quite simply, having a larger batch size facilitates learning the target grammar, and the larger the batch is, the less likely the learner is to make errors. To see this, consider again the nature of the ambiguity of TP-headedness in Swiss German. As discussed in previous sections, all matrix clauses are ambiguous for T-initial/final, as are all embedded clauses. However, in conjunction

[33] This approach will not work in cases of diachronic change, in which the 2 sentences of input in the batch result from different grammars (i.e. different sets of parameter values). As speculation, we might suppose that at some point the learner could avail itself of splitting this batch into smaller batches, each sub-batch consistent with a single grammar.

with each other the two kinds of input are often unambiguous for T-final. For example, while both *SAuxV* and *SV[CompSVAux]* are individually ambiguous, only a T-final grammar can account for both of them. Suppose, then, that the learner stored these 2 clauses as part of a batch. Such a batch would provide unambiguous evidence for T-final. The larger the batch is, the more likely it is that the learner will be faced with unambiguous evidence for T-final. This means that the frequency of unambiguous evidence will also increase. And the more unambiguous evidence there is, the more rapidly the learner can converge on the adult grammar. Conversely, the proportion of ambiguous evidence will decrease. I have shown that ambiguous evidence plays a crucial role in understanding learners' errors. Thus, as the frequency of ambiguous evidence decreases, the likelihood of learners mis-setting a parameter also decreases. In sum, learning with larger batch sizes has a direct impact on the learning results. If one were to pursue the idea that learners did learn from larger batches, one would need to have greater confidence in the rapidity of acquisition. It is not obvious that one could be so confident, especially given the fact that child errors reflecting non-target parameter values are attested. Nevertheless, the extent to which learning from larger batches can accurately reflect child acquisition is ultimately an empirical question that should be investigated more fully.

In this section I presented a broader discussion of how input relates to learning in the model I have proposed. The central claim is that input frequency plays a key role in parameter setting. I also explored various ways in which the relationship between the input and learning might be modified, such as whether the model might filter the input it learns from (the question of uptake), and how much input the model stores in order to learn from at any given point in time (the question of batch learning). These modifications are compatible with the basic architecture of the model, but depending on how they are actually implemented may or may not yield results that are consistent with the results I have reported in this chapter. Above all, given what is currently known about the acquisition process, these modifications should be viewed as speculative departures from the null hypothesis (i.e. the basic implementation of the model used here).

9. Summary

In this chapter I have addressed the learning puzzle presented by the acquisition data of production errors concerning verb placement by Swiss German children. I have presented a learning model that relies on the parameter interaction found in ambiguous evidence to initially push the learner toward a non-target parameter setting. The probabilistic learner thus accounts for these production errors in a systematic way. Moreover, the learner captures the variable nature of these errors in a later stage of the children's developmental course. In the following chapter we will see another application of learning from ambiguous evidence via parameter interaction, namely variability across speakers in setting the same parameter.

Appendix 1: Swiss German input types and corresponding compatible grammars

Table 3.5 shows all 32 grammars from the 5 parameter space and which types of input from (31) that they are compatible with.

Table 3.5

Grammars	Input Types
(1) [+VT, +TC, C-init, T-init, V-init]	SV, SVO, SAuxV, SV[Comp. SV], XVS, XVSO, XAuxSV, OVS, OAuxSV
(2) [+VT, +TC, C-init, T-init, V-fin]	SV, SVO, SAuxV, SAuxOV, SV[Comp. SV], XVS, XVSO, XAuxSV, XAuxSOV, OVS, OAuxSV
(3) [+VT, +TC, C-init, T-fin, V-init]	SV, SVO, SAuxV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux], XVS, XVSO, XAuxSV, OVS, OAuxSV
(4) [+VT, +TC, C-init, T-fin, V-fin]	SV, SVO, SAuxV, SAuxOV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux], SV[Comp. SOVAux], XVS, XVSO, XAuxSV, XAuxSOV, OVS, OAuxSV
(5) [+VT, +TC, C-fin, T-fin, V-fin]	SV
(6) [+VT, +TC, C-fin, T-fin, V-init]	SV
(7) [+VT, +TC, C-fin, T-init, V-fin]	SV
(8) [+VT, +TC, C-fin, T-init, V-init]	SV
(9) [+VT, –TC, C-init, T-init, V-init]	SV, SVO, SAuxV, SV[Comp. SV], XVS, XVSO, XAuxSV, OVS, OAuxSV
(10) [+VT, –TC, C-init, T-init, V-fin]	SV, SVO, SAuxV, SAuxOV, SV[Comp. SV], XVS, XVSO, XAuxSV, XAuxSOV, OVS, OAuxSV
(11) [+VT, –TC, C-init, T-fin, V-init]	SV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux]
(12) [+VT, –TC, C-init, T-fin, V-fin]	SV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux], SV[Comp. SOVAux]
(13) [+VT, –TC, C-fin, T-fin, V-fin]	SV
(14) [+VT, –TC, C-fin, T-fin, V-init]	SV
(15) [+VT, –TC, C-fin, T-init, V-fin]	SV, SVO, SAuxV, SAuxOV, XVS, XVSO, XAuxSV, XAuxSOV, OVS, OAuxSV
(16) [+VT, –TC, C-fin, T-init, V-init]	SV, SVO, SAuxV, XVS, XVSO, XAuxSV, OVS, OAuxSV
(17) [–VT, –TC, C-init, T-init, V-init]	SV, SVO, SAuxV, SV[Comp. SV], XAuxSV, OAuxSOV
(18) [–VT, –TC, C-init, T-init, V-fin]	SV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux], SV[Comp. SOVAux]
(19) [–VT, –TC, C-init, T-fin, V-init]	SV, SVO, SAuxV, SV[Comp. SV], XAuxSV, OAuxSOV
(20) [–VT, –TC, C-init, T-fin, V-fin]	SV, SV[Comp. SV], SV[Comp. SOV],

	SV[Comp. SVAux], SV[Comp. SOVAux]
(21) [–VT, –TC, C-fin, T-fin, V-fin]	SV
(22) [–VT, –TC, C-fin, T-fin, V-init]	SV, SVO, SAuxV, XAuxSV, OAuxSOV
(23) [–VT, –TC, C-fin, T-init, V-fin]	SV
(24) [–VT, –TC, C-fin, T-init, V-init]	SV, SVO, SAuxV, XAuxSV, OAuxSOV
(25) [–VT, +TC, C-init, T-init, V-init]	SV, SVO, SAuxV, SV[Comp. SV], XAuxSV, OAuxSOV
(26) [–VT, +TC, C-init, T-init, V-fin]	SV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux], SV[Comp. SOVAux]
(27) [–VT, +TC, C-init, T-fin, V-init]	SV, SVO, SAuxV, SV[Comp. SV], XAuxSV, OAuxSOV
(28) [–VT, +TC, C-init, T-fin, V-fin]	SV, SV[Comp. SV], SV[Comp. SOV], SV[Comp. SVAux], SV[Comp. SOVAux]
(29) [–VT, +TC, C-fin, T-fin, V-fin]	SV
(30) [–VT, +TC, C-fin, T-fin, V-init]	SV, SVO, SAuxV, XAuxSV, OAuxSOV
(31) [–VT, +TC, C-fin, T-init, V-fin]	SV
(32) [–VT, +TC, C-fin, T-init, V-init]	SV, SVO, SAuxV, XAuxSV, OAuxSOV

Chapter 4

Head-finality and Verb Movement in Korean: Modeling variability and non-variability across learners

1. Introduction

In this chapter I consider the following learning puzzle: what might the end-state of the learner's grammar look like when learning exclusively from ambiguous evidence? How will the learner set parameters, and will these parameter settings be consistent across learners of the same language? As in the previous chapter, we shall see that ambiguous evidence can be highly informative to the learner by means of parameter interaction. In this chapter I will address these learning questions with an empirical case study that looks at parameter setting in canonically verb-final languages. The discussion in this chapter is in principle relevant for all verb-final languages, but I will limit the discussion to Korean (and by extension to Japanese). This was done so as to focus on the rich experimental and acquisition work done in such recent studies on Korean as Han et al. (2007), the results of which dovetail with those of the model here. Additional experimental work begun in Han et al. (2008) has revealed similar results in Japanese. The crucial finding of the modeling results is two-fold: for some parameters the model can consistently learn the same value, whereas for other parameters there is variability as to which value the model learns.

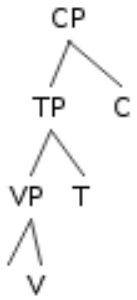
To give a general illustration of the learning puzzle, I will focus on the following example from Korean, which is given schematically in (1a).

(1) a. SOV
b. Chelswu-ka ppang-ul mek-ess-ta. (Hagstrom 2002: 211)
Chelswu-nom bread-acc eat-past-decl
'Chelswu ate the bread.'

In the discussion below, we will see that input such as this is full of ambiguities for the learner, from whether there is verb raising to the direction of heads and their complements. This ambiguity raises the question: what kind of grammar will the learner acquire?

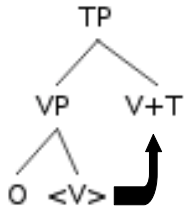
Ambiguity concerning verb raising is amply discussed in GB/Minimalist literature. A verb-final word order is commonly assumed (sometimes implicitly) to have consistently head-final projections along the clausal spine, as in (2), but tense morphology could appear on the verb in multiple ways.

(2)



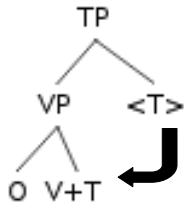
This could simply be a result of V-to-T movement, as shown schematically in (3). Angled brackets are used to indicate syntactic positions where a constituent appears during the course of a derivation but where it is not pronounced.

(3) *Verb raising*
a. [+V-to-T, V-final, T-final]
b.



Another possibility is that the verb can remain in-situ and either (a) the verbal suffixes lower via some rule of affix lowering/hopping (e.g. Chomsky 1957, 1981), or (b) the verb and its suffixes combine via some post-syntactic morphological operation such as Marantz's 1988 Morphological Merger (cf. other post-syntactic operations in more recent work such as Embick and Noyer 2001). These different options are presented in (4), in which both VP and TP are head-final, but there is no V-to-T movement. I assume that if there is no verb movement, then some operation such as (4b) or (4b') must take place in order for the tense affix to be part of the verbal complex (cf. Lasnik's 1981 'stranded affix' filter). Such an operation, then, would simply be a reflex of a non-verb raising parameter setting.

(4) *Non-verb raising*
a. [–V-to-T, V-final, T-final]
b. Lowering

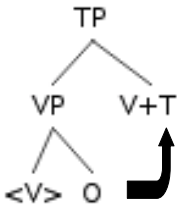


b'. Morphological Merger
$[_{TP}\ [_{VP}\ O\ V\ ]\ T\ ] \rightarrow [\ [\ O\ ]\ [\ V+T\ ]\ ]$

Various researchers have made claims *for* verb raising in Korean/Japanese (e.g. Otani and Whitman 1991, Koizumi 2000, and Choi 1999) and *against* verb raising (e.g. Yoon 1994). Faced with canonical data such as (1), which is ambiguous as shown in (3) and (4), these strands of research have drawn on a diverse set of empirical phenomena (e.g. null object constructions and coordination constructions to name a few) that are claimed to provide insight on the issue of verb raising. However, upon closer examination, the data either have no bearing on the issue of verb raising or are in principle compatible with either a raising or a non-raising analysis. This is discussed concisely in Han et al.'s (2007) overview of the literature, to which can be added Takano (2002) for a non-verb raising alternative to Koizumi's (2000) argument for verb raising. There is certainly no consensus on the issue of raising, and short of a clear empirical argument to the contrary, I will assume that either a raising or non-raising grammar is a viable target for a child learning Korean.

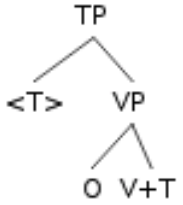
That there is ambiguity with respect to verb raising is well known. In fact, following along the lines of Han et al. (2007), I will claim that learners can acquire either a verb raising or non-verb raising grammar. It is perhaps surprising to note, though, that data such as (1) are also ambiguous with respect to headedness. This input is compatible with either VP or TP being head-initial. VP can be head-initial if the verb vacates the VP and moves to a head-final TP, as in (5).

(5) *Verb Raising*
a. [+V-to-T, V-initial, T-final]
b.



Further, TP can be head-initial if the verb remains in-situ with some affix lowering/morphological merger:

(6) *Non-verb-raising*
a. [–V-to-T, V-final, T-initial]
b.



I note that recent work such as Biberauer et al. (2014) claims that the structure in (5) is unattested cross-linguistically. Nevertheless, I will assume that a grammar such as (5) is

not automatically ruled out for the learner. As I discuss in Section 7, by crucially allowing the learner to consider combinations of parameter values such as (5a), it is possible to model variability with respect to verb movement. Ultimately, the question is whether the learner will end up acquiring a grammar such as that in (5). In Section 7, I illustrate some of the difficulty involved in trying to address this question. I return to this question in the more general discussion in Chapter 6 with a proposal for how it can be addressed in future research. As regards Korean, though, which is of more immediate interest here, we will see that the model can always learn a grammar that falls within the range of grammars that Biberauer et al. claim are attested.

In Korean linguistics, there is a tradition of referring to tense and mood morphology on the verb as phrasal affixes, here phrasal suffixes (Yoon 1994; Park 1994). For example, tense morphology instantiates a T-head that takes as its complement a VP. As a phrasal suffix, tense morphology affixes to the constituent it subcategorizes for (i.e. the VP), and as the VP is assumed to be verb-final, it will suffix to the verb. However even under this approach, the precedence relation of the VP relative to T is underdetermined: so long as T is a suffix, TP can be head-initial or head-final, and the verbal complex will appear in the desired form because T will affix to a head-final VP. Thus the phrasal affix approach does not address some of the basic ambiguity under discussion, and I will not consider it any further. Moreover, this approach presupposes that the verb does not move because the phrasal affix T is attaching to a VP that contains an in-situ verb. In contrast, I claim that there is variability across learners regarding verb movement.

Thus the input in (1) underdetermines the basic structural analysis of the clause. As will be discussed further, this is a general property of not just (1), but of all (or what appears to be nearly all) the input that a learner of a verb-final language such as Korean receives. The claim, then, is that the learner is faced with insufficient unambiguous evidence for parameter setting. In Chapter 1 I introduced the subset learning scenario, in which there is also insufficient unambiguous evidence, and discussed how learning from implicit negative evidence can address this issue. This will be illustrated in more detail in the case of causatives in Chapter 5. In the case of causatives, there are competing analyses involving varying degrees of structural complexity that result in subset/superset languages. An important consideration in this chapter is that none of the competing analyses in (3)-(6) generates a language that is obviously in any subset/superset relationship with any of the others. Thus the case of Korean does not lead to any considerations of the subset learning scenario and learning from implicit negative evidence.

Given the kind of ambiguity I have illustrated in (3)-(6) and the overall ambiguity of the input corpus, the puzzle is what do we expect as a learning outcome? Will learners favor some parameter settings over others? Will there be significant variation of grammars across or even within speakers? The null hypothesis is that learners of the same language will arrive at a grammar with the same parameter values. But what are the

predictions and results of the learning model? I have already alluded to variability across speakers concerning verb movement. Given modeling results for variability, is there additional evidence to support the existence of this variability? And what about other parameters? Is it possible for all speakers to learn the same settings for other parameter values?

The goal of this chapter is to propose a proof-of-concept learning model that shines a light on a previously undiscussed way of learning that helps provide an answer to these modeling questions. This model takes advantage of parameter interaction (which allowed us to model learner errors in Chapter 3) to systematically learn certain parameter settings. Although multiple grammars are compatible with all the input, we will see that because of parameter interaction, a majority of these grammars are V-final, and a majority of these grammars are T-final. As discussed in Chapter 1, a probabilistic learner can make use of this observation: the most probable and thus the most compatible grammar is one that is both V-final and T-final. Thus despite the pervasive ambiguity in the input, namely an entire corpus that is ambiguous for head-directionality parameters, learners of Korean or other verb-final languages are predicted to systematically learn on certain parameter settings.

To be clear, the goal of this chapter is not to show that the model will learn that every single projection along the clausal spine is head-final in Korean. Given the pervasive ambiguity of the input, such a result is not predicted in the current implementation of the model if there is a very large number of syntactic projections. In this chapter I will consider learning scenarios with only a relatively small number of projections, the values of which are all consistently learned to be head-final. Suppose in fact that the child does acquire a grammar with a much larger set of projections. What is predicted is that the model will consistently learn that the *same* subset of projections is head-final. These projections are those that are structurally closer to the lexical verb (or auxiliary). Again, given that this result is based entirely on ambiguous evidence, I take this to be a rather striking outcome.

An extension of this approach is to observe that parameter interaction does not always determine which grammar the model learns. Given insufficient parameter interaction, we expect either of a particular parameter's values to be learned as a parameter setting. This is what we see in the case of V-to-T movement in the example from above. As will be discussed, there is a symmetry between having and not having V-to-T movement: there is an equal number of input-compatible grammars with either parameter value. The expectation, then, is that the learner could adopt either parameter setting. The ambiguity of the input leads to a situation of variability across learners: some learners have verb raising, whereas others do not. Again as discussed in Chapter 1, variability in grammars is a phenomenon that can be readily modeled with a probabilistic learner; indeed this kind of stable variation is something the model predicts given a sufficient lack of parameter interaction. The proof-of-concept model provides an

illustration of what circumstances are necessary for this variation to arise in this probabilistic framework.

As I have mentioned, the null hypothesis for learners of the same language is that they do not have grammatical variability. Evidence against this comes from the modeling results concerning verb movement. Is there additional support for the claim that there is variability? Possible empirical support for this variation in Korean comes from the careful experimental work in Han et al. (2007). They show that both adult and child populations split with respect to the scope of negation, and in their analysis they tie this split to variability in verb raising. Han et al. claim that this variability arises due to the overall ambiguity of Korean, and that there is insufficient input regarding the scope of negation for the average learner to rely on it to determine the position of the verb. Given some additional complications regarding their analysis and the model assumed here, it is an open question as to whether the proof-of-concept modeling of variability in verb raising is actually what is attested in Korean speakers. However, the model can serve as an instructive illustration of how this variation could in principle be modeled in Korean. Further, the evidence in Han et al. is suggestive that the model's results concerning grammatical variability are on the right track. I will offer some remarks concerning these complications, including a slight modification of Han et al.'s analysis, that attempt to unify the results of the model with those of Han et al.

The availability of multiple analyses that are compatible with the input relates to a methodological point from Chapter 3. There I discussed how we can use the results of the learning model as evidence when there are competing syntactic analyses. As mentioned, the results of the simulations reported in this chapter show that the model sometimes learns a verb raising grammar and sometime learns a non-verb raising grammar. As regards the debate on whether there is verb raising in Korean, the results lead to the following claim: an analysis that categorically claims that there is or that there is not raising in Korean is too strong. Rather, the results support the claim that multiple analyses are possible, and that there is grammatical variability across learners as to which analysis they adopt. Again, this claim finds support in the results of Han et al. (2007).

The findings in this chapter also relate to the relation between inflectional material and verb raising. In Chapter 3 I noted that according to Bobaljik's (2000, 2001) theory of the Rich Agreement Hypothesis, if there are at least two pieces of verbal morphology (each with its own syntactic projection) on the finite verb, then the verb must raise to T. The general approach I have pursued here is that inflectional material does not provide any evidence to the learner for or against verb movement. In Section 3.2 we will see examples where it is certainly plausible to treat the multiple suffixes on the verb as instantiating different pieces of inflectional material. If we treat these affixes as indeed being inflectional material, then according to Bobaljik there must be verb movement. However, if the modeling results here and the experimental results of Han et al. (2007) are on the right track, then these examples are still ambiguous with respect to verb

movement to T. The evidence for this grammatical variability can thus be taken as evidence against the claim that rich inflectional material entails verb movement.

The structure of this chapter is as follows. In Section 2 I review the core learning mechanism of the model and provide a more complete discussion of the learning scenario for SOV input, which was introduced in examples (3)-(6). This discussion lays out in more detail the basic logic of how parameter interaction can push the learner to favor certain parameter settings, whereas insufficient parameter interaction does not. For expository purposes I illustrate this with a simple clausal structure in Section 2 that involves only three parameters. This serves the purpose of illustrating schematically how the model arrives at variability for verb movement while simultaneously learning head-finality. Results of running the model with the 3-parameter hypothesis space are given at the end of Section 2. The logic of this schematic version of the model is more general and can be applied to a more enriched syntax of Korean. In Section 3 I expand the scope of the discussion by incorporating further syntactic structure and by considering how to enlarge the input corpus to make it more representative of a verb-final language such as Korean. Again, a head-final syntax and variability for verb movement are expected in this model of Korean. This variability dovetails with the variability reported in Han et al.'s (2007) experimental work, which I discuss in Section 4. Han et al.'s analysis of Korean raises some potential complications that emerge when we consider an enriched picture of Korean syntax. I address these complications at the end of Section 4 and show how the learning model is still amenable to the basic framework assumed by Han et al. Results of running the model's simulation of Korean are then given in Section 5. I consider several other learning models in Section 6 and look at what role parameter interaction plays in learning from ambiguous input in these models. Finally, in Section 7 I discuss more generally the idea of constraining the model in light of putative language universals and illustrate some of the challenges such considerations pose for future research by focusing on the example of variability in Korean.

2. Modeling the effects of parameter interaction: The core example

This section takes up the learning puzzle that is presented in examples (3)-(6) and presents a fuller discussion of how ambiguous input can drive a probabilistic learner toward certain parameter settings via parameter interaction. To illustrate the basic logic of the learning method, I will use a simplified learning scenario that focuses on learning from SOV input and in which there is a minimal number of syntactic projections along the clausal spine. We will see in Section 3 that the results of this discussion can be naturally extended to an enriched corpus with a more fleshed out syntactic structure that more accurately simulates Korean. In fact, as will be discussed in Section 3, SOV input is a core input type that allows for learning with respect to headedness. This section thus serves as a schematic illustration of how to model variability for verb movement while

systematically learning the same values for various parameters concerning head-complement order. Results are given in Section 2.2.

2.1 A schematic version of the model: Learning in a 3-parameter hypothesis space

Before looking directly at learning from SOV input in the schematic version of the model, I first review some core components of the model, including the hypothesis space and the general learning mechanism, which was introduced in Chapter 2. For a hypothesis space in the schematic version of the model, I will use simple syntactic structures involving two projections along the clausal spine in order to illustrate the logic of the proposal. Let us consider the case where there are 3 binary parameters: 2 parameters for head-directionality, and 1 parameter for verb movement. These parameters are given below.

(7) *TP-headedness parameter*
a. T-in(itial): T linearly precedes its sister complement
b. T-fin(al): T linearly follows its sister complement

(8) *VP-headedness parameter*
a. V-in(itial): V linearly precedes its sister complement
b. V-fin(al): V linearly follows its sister complement

(9) *V-to-T movement parameter*
a. [+VT]: finite V moves to T if no free morpheme in T blocks it
b. [–VT]: finite V does not move to T

The values of these parameters can be combined in a total of 8 grammars, although not all of them are compatible with SOV input. An important consideration here is that no complete set of sentences that could be generated by any one of these grammars is a proper subset of those of any other grammar. Thus any consideration of the 'subset problem' is not applicable here. Accordingly, as in Chapter 3, I it is possible to use the simplified version of the model that was introduced in Chapter 2, and I do so in this chapter as well. As in the simulation of Swiss German, there are no restrictions in the hypothesis space concerning the logically possible combinations of parameter values. This results in a rich hypothesis space containing some grammars that recent work such as Biberauer et al. (2014) claims are not attested cross-linguistically. An example of such an unattested grammar would be one in which TP is head-final and VP is head-initial. I return to the question of how such putative language universals relate to the shape of the hypothesis space of the learning model in the discussion in Chapter 6, which follows the case studies in Chapters 3-5. For the time being it is sufficient to note that when we look at the results of the simulation below, the model can always learn a grammar that Biberauer et al. claim is attested. Before looking at the 8 logically possible grammars and

seeing which ones are compatible with the input, I first review the general mechanics of learning from input.

I use the simplified version of the model introduced in Chapter 2. Learning in this model is driven by each token of input the learner receives. Upon encountering a token of input, the model samples a parameter value from the probability distribution of each parameter to form a possible grammar, or vector of parameter values. Initially, all parameter values are assumed to have equal priors. The sampled values are then compared with the input. If the sampled grammar is compatible with the input, then the probabilistic weights of those values are increased. This is done so as to maximize the likelihood of sampling such a compatible grammar in the future given the same input. In this respect, the model is similar to Yang's learning model (cf. Section 6.3). More technical discussion of the priors and update procedure can be found with the results in Section 2.2. In terms of the fully generative model from Chapter 2 (which outputs sentences that are compared against the input), if output matching the input could possibly be generated with the parameter values sampled, then those values are reinforced.

As in Chapter 3, for the sake of expediently illustrating the proof-of-concept model, this implementation of the model involves a simplification with respect to generating output. Unlike the generative of the model introduced in Chapter 2, which makes use of a more fleshed-out context-free grammar so as to generate output strings (which are then compared against the input), the model here only checks compatibility between sampled parameter values and the input. For more discussion of this simplification, see Chapter 2. I thus abstract away from a number of syntactic details that would be involved in generating output sentences that match the input in the schematic corpus. For example, the 3 parameters in the model do not represent how modifiers are represented in the syntax; a fully generative model that outputs full sentences would need to take this into consideration.[1] As discussed in Chapter 2, this simplification allows for a more expedient illustration of how the model learns. I mentioned in the introduction to this chapter how none of the grammars in this kind of hypothesis space result in any languages that are in a subset/superset relation with each other. A fully generative model will be crucial in Chapter 5 to learn the grammar of a subset language from implicit negative evidence. No such subset learning scenario is under consideration here. All things being equal, adding an enriched probabilistic context-free grammar, with further choices that do not interact with these three parameters, should not affect the overall learning course of the model presented here. Thus it is possible to use the simplified version of the model from Chapter 2.

Given this framework for learning, let us now consider how parameter interaction favors some parameter settings by returning to the 8 grammars that are logically possible

[1] I have also abstracted away from parameters concerning the position of nominal arguments. For more discussion on such parameters, see Section 4 and note 11.

in the 3-parameter space introduced above. Limiting ourselves to SOV input, (10) and (11) show that only 4 of the grammars are compatible with the input. This assumes that all specifiers are on the left and abstracts away from the position of the subject. Thus if the verb raises to T, the subject is assumed to be in SpecTP. Crucially, the object is assumed to be merged as a sister of the verb and to remain in that position; I return to the significance of this point in Section 4.2. Further, I assume that a [–VT] setting necessitates some operation such as affix lowering to join the tense affixal morphology to the verbal complex. This can be thought of as a universal constraint on the well-formedness of a grammar.

(10) *Grammars compatible with SOV*
a. [+VT, *T-fin*, V-in] c. [–VT, T-in, V-fin]
b. [+VT, *T-fin*, V-fin] d. [–VT, *T-fin*, V-fin]

(11) *Grammars incompatible with SOV*
a. [+VT, T-in, V-in] b. [–VT, T-in, V-in]
c. [+VT, T-in, V-fin] d. [–VT, T-fin, V-in]

In (10), the input is ambiguous for all three parameters when taken individually; further, the entire corpus of input (here just SOV) is fully compatible with the four competing hypotheses in (10). Nevertheless, the key observation is that the vast majority of the compatible grammars are *T-final* (75%) and V-final (75%). This is a direct consequence of parameter interaction: [–VT] with V-initial is incompatible with SOV, as is [+VT] with T-initial.

In other words, the core insight is that given (10), the probabilistic learning model is most likely to sample a T-final grammar and a V-final grammar. Parameter interaction has those two values show up in the majority of compatible grammars, and majority rules here. The model's updating procedure then attempts to maximize the likelihood choosing such input-compatible parameter values in the future. Therefore, given sufficient SOV input the learner will be pushed in a principled way toward a [T-fin, V-fin] grammar because such a grammar is, in a sense, the most compatible grammar. If we start with equal weights for all parameter values, when running this model the learner will systematically learn grammar (10b) or (10d).

Thus even when the model is presented with fully ambiguous input, parameter interaction can consistently drive the learner toward certain parameter values. In the case at hand, these values are V-final and T-final, resulting in a head-final grammar. In contrast, no parameter interaction in (10) constrains whether the model learns [+VT] or [–VT]. So long as the learner is not too conservative or tentative, the model then predicts two populations of learners for this kind of input (cf. Chapter 2 and discussion in Pearl 2007): those with and those without V-to-T movement. A very conservative learner might perpetually vacillate between the [+VT] grammar in (10b) or the [–VT] grammar in (10d) because both grammars are equally good at matching the input data. A less

conservative learner will choose one of the grammars, say [+VT], and begin to reinforce that choice by gradually shifting more and more probabilistic weight to it. As there is insufficient parameter interaction to favor the other grammar, nothing will cause the learner to shift away from an analysis that converges on this [+VT] grammar. But precisely because that early choice of [+VT] could just as likely have been [–VT], the same logic holds for learners to converge on a [–VT] in approximately the same proportion as those converging on a [+VT] grammar. The experimental work of Han et al. (2007) to be discussed in Section 4 shows that such less conservative learners in principle exist. Assuming such a learner for verb raising, the model thus illustrates how stable variation can arise across speakers in a population.

2.2 Results of the 3-parameter model

In this section I report results showing how the model learns from SOV input in the 3-parameter space described above in (7)-(9). In anticipation of the enriched corpus for the simulation of Korean in Section 3, I introduce here a second kind of input to the schematic corpus: intransitive SV strings.

SV input is highly uninformative to the learner in the 3-parameter model. As there are no parameters in the 3-parameter hypothesis space that explicitly concern the position of the subject, the model accounts for the subject 'for free' as it were. All 8 grammars in the 3-parameter hypothesis space are compatible with SV input so long as we assume that the subject has raised to SpecTP when the verb raises to T. Thus this input is fully ambiguous with respect to head-directionality, with neither head-initial nor head-final being favored for any parameter setting. Further, as no head-initial or head-final grammar is punished, we can maintain symmetry between raising and non-raising grammars: 4 [+VT] grammars are compatible, and 4 [–VT] are as well. However even if a large proportion of the input corpus is SV, so long as there is sufficient SOV input (which is informative for head-finality), the model can still consistently learn a grammar that is head-final for VP and TP. As we will see immediately below, it is possible to calculate that approximately 25% of the input is SOV in this simulation of Korean, an ample proportion for the model to learn from.

As a rough approximation of the distribution of SV and SOV strings, I present the frequencies in (12) at which the input types are presented to the learner. Input to the model is sampled randomly and presented to the learner item by item. The probability distribution over input types that the input is sampled from is based on averages of child directed speech to children learning Korean. The probability distribution of input types is simply a multinomial whose values are arrived at in the following two ways. The probability of an input type could simply be its average frequency rate in a corpus. Or the probability of an input could be the joint probability of certain schematic elements co-occurring; this is calculated by using the probabilities (i.e. the average frequency rates) of the schematic elements in question, assuming they are independent of each other. For

example, according to Fukuda and Choi (2009) approximately half of all verbs spoken to children in the sample are transitive, and according to Kim (2000: 345) on average there is object drop with approximately half of all transitive verbs in child directed Korean. If half of all clauses in the corpus are transitive, and if half of clauses contain an overt object, then the likelihood of a clause that is transitive with an overt object (i.e. is SOV) is .25.

(12) *Input type and probability*
a. SV $p = .75$
b. SOV $p = .25$

We can assess the success of the model by looking at the weights of different parameter values, averaged across different runs of the model. A strong weight for a particular parameter value can be taken to indicate that the model learns that value. Further, each run of the model can represent a different learner in the population.

Thus given the discussion in Section 2.1, we expect the following results. We expect a point in the learning process at which the averaged weights across all runs will be near 100% for V-final and T-final (13a). This is because in any given run the model is expected to have a point in the learning process at which there will be weights near 100% for the head-final parameter settings (14a). In contrast, at this point, because of variability with respect to verb movement, we do not expect uniformity across runs in the weight for V-to-T movement (13b). Variability will instead be reflected with some proportion of runs that averages near 100% for [+VT] with the remainder of runs averaging near 100% for [–VT] (14b). The variability predicted here thus runs counter to the null hypothesis, according to which there is no grammatical variability.

(13) *Predicted weights of parameter settings (averages for multiple runs)*
a. Weights near 100% for {V-fin, T-fin}
b. Weights not near either 100% or 0% { [+VT], [–VT] }

(14) *Predicted weights of parameter settings for a single run*
a. Weights near 100% for {V-fin, T-fin}
b. Weights near 100% for { [+VT] or [–VT] }

I have just described the input corpus for the model and the expected results. I now describe the prior probabilities and the procedure used to update these priors with posterior probabilities. The model is run with a program written in the Church programming language (Goodman et al. 2008). I used the *bher* implementation of Church.

As the null hypothesis, I assumed all parameter values are initially weighted equally and thus used equal priors. Each parameter will represented with its own dirichlet distribution, with initial pseudo-count values of 1 for all parameter values. The pseudo-

counts are the parameter weights and represent the learner's expectations regarding the shape of the adult grammar. The prior pseudo-count values are thus equal, and are low, which represents a learner with weak initial expectations. With all the input to the model being ambiguous, it is important not to have the initial weights be too great for parameter values. Given priors with stronger expectations (e.g. pseudo-count totals of (100, 100)), it will be harder for the model to be pushed strongly in one direction over another, especially in cases where there is symmetry between a parameter's two values. We can capture the effect of a learner that is not too conservative (cf. discussion above) by using priors with weak expectations, or parameter weights that are initially relatively small. Each parameter's dirichlet distribution uses its weights to generate a probability for a given parameter value. These probabilities are then used to sample a grammar for a given token of input. For example, the parameter for VP-headedness will initially have a dirichlet distribution of *dir*(1, 1), where the first pseudo-count total corresponds to V-initial, and the second to V-final. For more discussion of the dirichlet distribution, I refer the reader to Section 2.4 in Chapter 2.

The update procedure is as follows and applies equally to all parameters. For each token of input, the model will sample from the parameter weights so as to select 10 (potentially non-distinct) grammars that are compatible with the input. Each of these sampling procedures can be called a chew. These 10 chews represent what the model learns for each token of input. This can be expressed in terms of pseudo-counts. If only one grammar G_1 is selected in all 10 chews, the pseudo-counts of all the parameter values that comprise that grammar will be increased by 1. If G_1 is V-final after the first token of input, then in the example of VP-headedness, the updated pseudo-counts will be the dirichlet distribution of *dir*(.5, 1.5). This means that the adjustment to the weight of any parameter value being used after a single chew is 0.1. If after the first token of input some input-compatible grammars using V-final were sampled 7 times, while input-compatible V-initial grammars were sampled 3 times, then the updated dirichlet distribution for VP-headedness would be *dir*(.8, 1.2). This process iterates after each subsequent token of input, with the pseudo-count values increasing accordingly. However, to expediently see the proof-of-concept model shift most of the probability mass onto one parameter value over the other (and thus learn the parameter setting with a heavy weighting) without the pseudo-count values becoming very large, I have chosen to normalize the pseudo-count total of each parameter after a certain amount of input. Again, to model learners that are not too conservative, for each parameter I normalized the sum of the pseudo-count values to 10 after every 20 tokens of input. This means that the sum of the two pseudo-count values for, say, VP-headedness would equal 10 after 40 tokens of input. I note that the normalization process can sometimes result in the weights becoming very small. The pseudo-count values for the dirichlet distribution must be greater than 0, but if the value becomes too low, the model will treat it as if it is 0, resulting in a domain error. To avoid

this, a parameter's weights were adjusted to 0.02 and 9.98 if after being normalized, they were below 0.02 or greater than 9.98.

Results of running the model 15 times are given below in Table 4.1. These results show the average weight for opposing parameter values after running the model for an average of approximately 1,850 tokens of input. These averages are based on what proportion a parameter value's pseudo-count total is out of 10 (which is the sum of the pseudo-counts of all of a given parameter's values). A high average represents a point in the learner's development at which the model has learned the corresponding parameter value.

The results show high averages near 100% across all runs of the model for V-final and T-final. Depending on the run of the model, though, we have different results for verb raising. On some runs, one value will have a high average, while on other runs that value will have a low average. These results confirm the predictions in (13) and (14). For all runs of the model there is a point in the learning process at which the learner adopts a consistently head-final grammar. However, at that point of development, the grammar varies as to whether it is a raising or a non-raising grammar: the model has a non-zero probability of learning either. A binomial test was run under the hypothesis that there is a .5 probability of learning either a raising or non-raising grammar. Under such a null hypothesis, a success rate of 4/15 for learning a raising grammar has a two-tailed p-value of .1185. This means that if the population of Korean speakers is split evenly between verb raising and non-verb raising grammars, then we have a fairly reasonable likelihood (around 12%) of seeing 4 or fewer verb-raisers or 11 or more non-verb-raisers in a sample of 15 individuals.

Table 4.1 Average proportions of weights for parameter values
(Average of approximately 1,850 tokens of input per run)

Parameter Value	**Proportion**	**# Runs**
[V-fin]	.9777	15
[V-init]	.0233	
[T-fin]	.9655	
[T-init]	.0345	
[–VT]	.9442	11
[+VT]	.9015	4

Further, a look at the developmental course of the model reveals that more generally the model rarely assigns a strong weight to a head-initial value. In general, the weights for head-initial parameter values go steadily down. On only one run does a weight for a head-initial value rise above 75%. On this run the model initially strongly favors T-initial, but then the weight for T-initial drops, and the model strongly favors T-final. There is thus almost no variability in parameter setting throughout the course of

learning. The model is consistently learning head-final values for VP and TP but varies as to whether the grammar has or doesn't have verb raising.

Thus we have some evidence against the null hypothesis that there is no grammatical variability across speakers. Additionally, this result takes us a step closer to understanding the experimental results in Han et al. (2007), which will be discussed in more detail in Section 4. Based on results that vary across speakers, Han et al. claim that some speakers of Korean have verb movement, whereas others do not. The schematic model here is a first pass at trying to model this kind of variability in a principled way.

With this schematic version of the model, we have now seen some answers to the questions that opened this chapter. Can the model learn certain parameter settings in a principled way given ambiguous input? Yes: provided there is sufficient parameter interaction, certain parameter settings are favored, and the model will learn grammars with those parameter settings. Do we also expect variation across speakers in learning certain parameters? Again the answer is yes, but this time the result depends on parameters not interacting, leaving multiple parameter settings as equally likely. The next question is how general the scenario described in (10) above is. In the following section I show that by expanding the size of the parameter space and the input in the corpus so as to give a rough approximation of Korean in a simplified form, we have the same learning dynamics. There is a basic tension between raising and non-raising grammars, whereas head-finality is consistently favored. Results of this simulation of Korean are given in Section 5.

3. Making the model more general: A simplified Korean

In this section I discuss how the logic of the learning scenario from the previous section generalizes to a larger parameter space with an expanded corpus. By expanding the scope of the model, we can see how the model fares with a kind of simplified Korean. The goal of expanding the model, of course, is not to create the most realistic simulation of Korean, but to show the general applicability of the proof-of-concept model (a) in the abstract; and (b) to learning conditions that approach those of a child learning Korean. The working hypothesis of such an approach is that further enrichment of the model approaching the complexity of actual Korean does not affect the general learning results reported here. Nevertheless, there are some potential complications, which I discuss in Section 4, before reporting the results of the model's simulation of Korean in Section 5.

In addition to presenting a more complete picture of Korean, there are two other important aspects of the expanded model. First, we will see that there is a more nuanced kind of parameter interaction, what I call *secondary parameter interaction*. This is when learning one parameter value has a more local conditioning effect on learning another value. We will see an example of this kind of interaction in this section, and I will return to another example of it in Section 5. Second, a more enriched model brings us closer to modeling the experimental results of Han et al. (2007). Recall that on the basis of

variable experimental results, Han et al. concluded that there was variability across speakers of Korean with respect to verb raising. We would like to see, then, whether variability obtains in the modeling results with a richer syntax that more closely approximates that of actual speakers. And indeed it does, as reported in the results in Section 5.

I consider two kinds of ways of expanding the model. The first is to expand the hypothesis space in Section 3.1; the second is to expand the input corpus in Section 3.2. With additional parameters in the hypothesis space, we still see the same basic kinds of ambiguity that have been illustrated so far in this chapter. Further, the investigation of additional types of input here leads to the hypothesis that virtually all the evidence in Korean is ambiguous along the lines discussed above. In Section 4, we will see that Han et al. (2007) propose that there is some unambiguous evidence (which involves negation), but Han et al.'s results suggest that this input is extremely rare. Based on this, I will follow them in assuming that this unambiguous evidence is sufficiently rare so as to not play a role in parameter setting, at least for the average learner. This leads to the strong claim that the parameters under consideration must be learned exclusively on the basis of ambiguous evidence. Thus Korean provides an empirical example for how to address on of the acquisition questions from Chapter 1: how to learn parameter setting, and how to account for grammatical variability given an input corpus with insufficient unambiguous evidence.

3.1 Expanding the hypothesis space

To begin, let us consider a more complex hypothesis space by adding an additional functional projection along the clausal spine between V and T. The exact identity of such an intervening head does not affect the general point of the discussion, but for the sake of concreteness, let us assume that it is a little *v*-head. With this additional structure we can add several parameters to the hypothesis space. The first is a head-directionality parameter for the *v*-head, and the second is a movement parameter for *v*-to-T movement. Third, we can modify the parameter involving head-movement of V so that it is now for V-to-*v* movement. The five parameters used in this expansion of the model are given below, with (7) and (8) repeated from above.

(7) *TP-headedness parameter*
  a. T-in(itial): T linearly precedes its sister complement
  b. T-fin(al): T linearly follows its sister complement

(8) *VP-headedness parameter*
  a. V-in(itial): V linearly precedes its sister complement
  b. V-fin(al): V linearly follows its sister complement

(15) *vP-headedness parameter*
a. *v*-in(itial): *v* linearly precedes its sister complement
b. *v*-fin(al): *v* linearly follows its sister complement

(16) *v-to-V movement parameter*
a. [+V-*v*]: finite V moves to *v* if no free morpheme in *v* blocks it
b. [–V-*v*]: finite V does not move to *v*

(17) *v-to-T movement parameter*
a. [+*v*-T]: *v* moves to T if no free morpheme in T blocks it
b. [–*v*-T]: *v* does not move to T

Assuming the Head Movement Constraint (Travis 1984) is operative, then with the five binary parameters above, there are 32 possible grammars in the hypothesis space of the learner. Again no complete set of sentences that could be generated by any one of these grammars is a proper subset of those of any other grammar (cf. Safir 1987; Atkinson 2001). And if the verb does not move to a higher head, then the affixal material of that higher head must join the verbal complex via some other operation (e.g. affix lowering).

I note that the parameter space provided by these parameters is comparable to that assumed in Han et al. (2007), both in the kinds and numbers of parameters. If we follow the analysis of Han et al. relatively closely, then we can use this expanded parameter space as the basis for modeling results that are roughly comparable to those of Han et al. (see Sections 4 and 5). Let us next consider which of these grammars is compatible with the input.

Having introduced the parameter space, let us now look at the grammars it generates to see which ones are compatible with the input. The conclusions from Section 2 extend naturally to the case of having only SOV input in the 5-parameter implementation of the model. Again, SOV input is ambiguous with respect to all 5 parameters, but some of the parameter values are more probable among the grammars compatible with the input. For each headedness parameter the vast majority of compatible grammars are head-final, whereas there is symmetry for the raising parameters. To see this in the more complex parameter space, it is perhaps convenient to discuss the grammar space as having grammars that are composed of two relations, one mediated by head-movement to *v*, and the other by head-movement to T.

When we consider the relation between V and *v*, we can use essentially the same logic that was used to discuss the relationship between V and T in (10) and (11) in Section 2. Simply replace T with *v* and the following is true: (a) if the verb remains in-situ, then V must be head-final; (b) if V raises to *v* (and does not raise to T), then *v* must be head-final; and (c) half the grammars compatible with the input have V-to-*v* raising, while half do not. This means that the majority of compatible grammars are V-final and *v*-final, whereas half the compatible grammars [+V-*v*], and half are [–V-*v*]. The full list of 16 input-compatible grammars is given in (18) below, where this observation can be

verified; cf. the tallies in (19). Note that whether or not the model learns a verb raising grammar can be determined based simply on whether it learns a [+V-*v*] setting or not. It is sufficient for the verb to have moved at all in order to be considered a verb raising grammar. Thus half the grammars in (18) have verb raising, and half do not.

Next, consider the relation between *v* and T; the same kind of argument repeats. If the verb does not raise from the VP, then all the overt morphology will appear in the VP, and there are no restrictions on whether T is head-initial or final or on whether *v* moves to T. However, of the 8 grammars that are [+V-*v*], then the following is true: (a) if the verb raises to T, then T must be head-final; (b) as stated above, if V raises to *v* and does not raise to T, then *v* must be head-final; and (c) half of these compatible [+V-*v*] have *v*-to-T raising, while half do not. This means that the majority of compatible grammars are *v*-final and T-final, whereas half the compatible grammars are [+*v*-T], and half are [–*v*-T] is equally probable. Interestingly, when we look at the full list of compatible grammars in (18), we see that only 25% of them involve raising the lexical verb all the way to T. Even though half the grammars are [+*v*-T], only one quarter of them are [+V-*v*] and [+*v*-T], raising the verb to higher projections being dependent on having learned to raise it to lower ones. This suggests there is additional variation across speakers based on how high they raise the verb: a smaller proportion of the population will raise the verb to T than will raise it to *v*.

(18) *Grammars compatible with SOV input (5-parameter hypothesis space)*

a. [+V-*v*, +*v*-T, T-fin, *v*-in, V-in]
b. [+V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
c. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-in]
d. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]

e. [+V-*v*, –*v*-T, T-in, *v*-fin, V-in]
f. [+V-*v*, –*v*-T, T-in, *v*-fin, V-fin]
g. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-in]
h. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]

i. [–V-*v*, +*v*-T, T-in, *v*-in, V-fin]
j. [–V-*v*, +*v*-T, T-in, *v*-fin, V-fin]
k. [–V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
l. [–V-*v,* +*v*-T, T-fin, *v*-fin, V-fin]

m. [–V-*v*, –*v*-T, T-in, *v*-in, V-fin]
n. [–V-*v*, –*v*-T, T-in, *v*-fin, V-fin]
o. [–V-*v*, –*v*-T, T-fin, *v*-in, V-fin]
p. [–V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]

(19) *Summary of grammars in (18)*

a. V-fin grammars: 75%
V-in grammars: 25%

b. *v*-fin grammars: .625%
v-in grammars: .375%

c. T-fin grammars: .625%
T-in grammars: .375%

d. +V-*v* grammars: 50%
–V-*v* grammars: 50%

e. +*v*-T grammars: 50%
–*v*-T grammars: 50%

f. [+V-*v*, +*v*-T] grammars: 25%

We see then that adding additional syntactic structure results in predictions for the model that are largely the same as before. Given sufficient SOV input, learners will systematically learn that each of the projections along the clausal spine is head-final because that is the most probable structural analysis. Also, there is symmetry between raising and non-raising grammars. One difference here is the possibility for further stable variation with respect to the height of verb raising. Still, we expect some proportion of learners to have verb raising.

Adding in even more syntactic structure does not change the logic of this approach. If we add an additional functional projection XP, then the model can learn that XP is head-final provided that the learner sufficiently considers the possibility that the verb raises to X. If the verb moves to X, then X must follow the object, which means that XP must be head-final.

There is an interesting wrinkle to this approach, though. We saw that the higher XP is in the structure, the less likely it is for the model to learn the verb moves to X; this is because moving the verb to X is dependent on having learned to move the verb through all the heads of lower projections.[2] Therefore, as the model is less likely to learn verb-raising to X, it is also less likely to learn that X is head-final. And the higher XP is in the structure, the less likely the model is to learn that (a) the verb moves to X; and (b) that XP is head-final. Nevertheless, this does not have an effect on what the model learns about projections that are lower down in the clause. If, for example, we added CP and

[2] This assumes that all head-movement parameters are formulated as being highly local, i.e. as involving movement from the head of a complement to the head that selects it. We could consider additional kinds of head-movement parameters, such as one involving V-to-C movement, as was discussed in Chapter 3 (note 10), which has the verb move cyclically through a number of intermediate projections. Preliminary investigation indicates that these kinds of parameters will change somewhat the predictions resulting from parameter interaction, but that there are still broad similarities with those reported in the main text. I will consider one example of this here, reserving a more careful investigation for future research. For example, suppose we augmented the 5-parameter hypothesis space above with parameters for CP-headedness, T-to-C movement, and V-to-C movement. With respect to verb movement, this new 8-parameter hypothesis maintains symmetry, and we expect variability with respect to learning a verb raising or non-verb raising grammar. As regards head-finality, CP will be heavily favored via parameter interaction to be head-final. Further, phrases lower in the structure will be favored to a lesser degree to be head-final, with phrases higher in the structure but below CP – for example TP – being favored least strongly to be head-final. Again we see that head-finality is favored in general, although the details are somewhat different from those presented in the main text.

parameters for C-headedness and T-to-C movement, the proportions in (19) would not change, although the proportion of C-final grammars that are compatible with the input would now be a smaller majority of only 56.25%. Thus the observations about learning head-finality concerning the phrases in (19), as well as a split in the population regarding verb movement – these observations continue to hold with the addition of further functional projections.

In sum, additional functional structure does not change what the model can learn from ambiguous input. It will systematically learn that certain, structurally lower projections, are consistently head-final. Further, it will sometimes learn that there is verb raising and sometimes learn that the verb remains in-situ.

3.2 Expanding the corpus

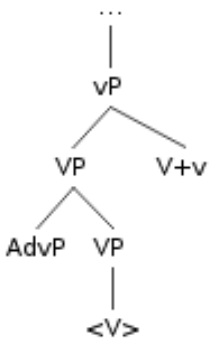
I turn now to an expansion of the corpus of schematic input, which will give us a rough approximation of a simplified Korean. There are two core kinds of input that we can use to build this corpus: input where V takes a complement as its sister, and input where the V does not take a complement. Consistent with the schematic model in Section 2, I will ultimately settle on a schematic corpus of just two types of input – SV and SOV – a coarse level of detail, but as I will discuss, one that actually covers the major patterns of evidence for learning the syntactic parameters at hand.

Let us now consider the two kind of core input for a schematic corpus, transitive and intransitive clauses. The prototypical input where the verb does not take a complement is intransitive SV strings. As discussed earlier, this kind of input is highly uninformative to the learner. All 32 grammars in the five-parameter hypothesis space are compatible with SV input so long as we assume that the subject has raised to (at least) SpecTP when the verb raises to T. Thus this input is fully ambiguous with respect to head-directionality, with neither head-initial nor head-final being favored for any parameter setting. Further, as no head-initial or head-final grammar is punished, we can maintain symmetry between raising and non-raising grammars: 16 [+V-*v*] grammars are compatible, and 16 [–V-*v*] are as well. However even if a large proportion of the input corpus is SV, so long as there is sufficient SOV input (which is informative for head-finality), the learner can still learn head-final settings for the relevant projections. As we will see below, it is possible to calculate that approximately 25% of the input is SOV in this simulation of Korean, an ample proportion for the model to learn from.

It is interesting to note that there is at least one kind of common intransitive data in Korean that at first glance appears to actually break the symmetry between raising and non-raising grammars in the hypothesis space we have been considering. Discussing this data also provides a helpful example of how the effects of learning from ambiguous evidence can be mitigated if an insufficient proportion of it occurs in the learner's input. This is data where a modifier intervenes between the subject and the verb. For VP-modifiers, for example, this can be schematized as S[$_{VP}$AdvV]. This kind of input is still

ambiguous as regards verb-raising and headedness for the same reasons that we have seen throughout. However, this kind of data will actually favor a non-raising grammar. To see this, let us consider the structure in (20) with VP-modification and focus on whether there is at least raising to *v*. The discussion here will focus on adverbial modification of VP, but similar generalizations also hold for adverbial modification of *v*P or TP.

(20)



When the verb raises to *v*, then the grammar must be *v*-final because the verb must follow the modifier. Raising to *v* punishes *v*-initial (eliminating 4 of the 32 grammars), and *v*-finality is favored for now familiar reasons. Similarly, if the verb raises to T through *v*, TP must be head-final. However, should the verb remain in-situ in the VP, then either a V-initial or V-final grammar is compatible in (20): with either grammar the verb will follow the modifier. As neither direction is punished with no verb raising, there are now more non-raising grammars that are compatible with (20):

(21) *Grammars compatible with [S [$_{VP}$ AdvV]] (5-parameter space)*
a. +V-*v* grammars: 8/24 = 33.33%
b. –V-*v* grammars: 16/24 = 66.66%

Given that [–V-*v*] is favored in (21), we might expect that the learner will always converge on a non-raising grammar. This is not in principle a problem for the model, but if we follow the discussion in Section 4 based on Han et al. (2007) that Korean speakers are split between raising and non-raising grammars, the model would not reflect this variability if it could never learn a raising grammar.

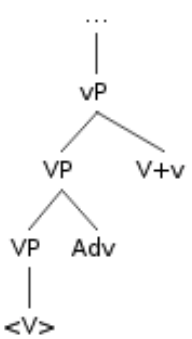
Although [–V-*v*] grammars are favored in (21), a verb raising grammar can still be learned. In particular, two ways of learning a verb raising grammar come to mind. The first is to simply note that the parameter interaction in (21) in favor of not raising can be overcome if the proportion of VP-modified input is sufficiently low. If the learner only rarely encounters input such as (20), then the learner's input is effectively neutral with respect to favoring raising or non-raising grammars. That is, the vast majority of the time there is symmetry between raising and non-raising grammars; as was discussed above, with such symmetry a learner that is not too conservative can learn either kind of

grammar. Indeed, when input such as (20) is added to the 5-parameter model with the schematic corpus presented later in (28) and the model is run, it is possible to learn the raising grammar (as well as the non-raising grammar) so long as the proportion of S[$_{VP}$AdvV] input is at least 3.5% or lower.

I note that this 3.5% reflects a frequency for VP-modification, not verbal modification. If *v*P and VP are equally likely to be modified (cf. a similar assumption made in Chapter 5), then the rate of verbal modification in the corpus would be at least 7%. Input with *v*P modification also favors not raising to *v*, although to a lesser degree: with such input, 12/28 grammars are raising grammars (42.85%), while 16/28 are non-raising ones (57.14%). Again, the impact such input has on learning a raising grammar is limited provided it occurs at a sufficiently low frequency. Future testing of the model can look at whether a raising grammar can be learned when the corpus is augmented with both types of input, both VP-modified and *v*P-modified strings.

There is another approach to the effect of parameter interaction we see in (21). The proportions in (21) are based on grammars that only have left-adjunction of modifiers. If we allow for the possibility of rightward adjunction, the push toward a non-raising grammar with this input is noticeably attenuated. This is because rightward adjunction of the modifier is only compatible with verb raising. We could represent the direction of adjunction with an additional parameter: there could be a parameter for leftward or rightward adjunction of verbal modifiers. Input of the form S[$_{VP}$AdvV] is ambiguous for leftward or rightward adjunction of the modifier. The 28 compatible grammars compatible with leftward adjunction have been summarized in (20); these grammars favor non-raising. However, rightward adjunction is possible so long as the verb raises past the site of adjunction. For example, in (20') the adverb is right-adjoined to VP, and the verb can only appear finally by moving at least as high as *v*.

(20')



Only grammars involving verb movement are compatible with rightward adjunction of VP-modifiers. Using the 5 basic parameters for headedness and verb raising, there are now 8 compatible grammars that have both verb raising and rightward adjunction. Added to the 12 verb raising grammars in (21), the total number of raising grammars compatible with S[$_{VP}$AdvV] is now 20; this exceeds the total of 16 for the non-raising grammars from (21b). Compared to (21), there is now an even slighter edge for raising (55% of

compatible grammars) as opposed to non-raising grammars (45%). Given a sufficiently small proportion of input with modifiers, though, we do not expect this new parameter interaction to result in learning only a raising grammar as has been made clear in the preceding discussion.

In fact, there is even further parameter interaction and more of a back-and-forth between favoring a raising or non-raising grammar. Although including the option of rightward adjunction results in raising being favored overall, leftward adjunction is favored to rightward adjunction (28 grammars to 8). And as we have seen, leftward adjunction favors non-raising. Thus we expect the learner to be pushed toward a grammar with leftward adjunction, which favors non-raisng, but we also expect leftward adjunction's favoring non-raising to be attenuated by the initial interaction that favors raising. In sum, the basic tension between raising and non-raising appears to persist in these more complex cases involving input with modifiers.

We now have a tentative range (of at least 3.5% of the corpus containing VP-modification) for how well the model performs with intransitive VP-modified input in the 5-parameter hypothesis space, and this can be compared to actual corpus frequencies of modification in future work. I have also discussed expanding the parameter set (e.g. direction of adjunction for modifiers) as another possibility for learning from this kind of input. This too can be investigated in subsequent research. The case of VP-modified input can thus show us the limitations of the model, but it is not in principle incompatible with modeling grammatical variability under the approach pursued here.[3] For the purpose of the proof-of-concept model, though, I will focus on the simpler case of adding just SV intransitive input to the corpus. This results, so far, in no change to the schematic corpus of {SV, SOV} strings. The kind of additional parameter interaction that we saw with leftward adjunction – i.e. given a push toward one parameter value (leftward adjunction), some other value (non-raising) is favored – is what I will call *secondary parameter interaction*. This kind of parameter interaction will be discussed further with the results in Section 5, where another example of secondary parameter interaction that favors non-raising will be identified.

I have now considered intransitive input. Although it presents some additional complexities, its overall profile is similar to the intransitive input we saw in the schematic version of the model in Section 2.

---

[3] Another type of input that is relevant here is SVO, which is also possible in Korean:

(i) Chelswu-ka mek-ess-ta sakwa-lul. (Choe 1987: 40)
Chelswu-nom eat-past-decl apple-acc
'Chelswu at an apple.'

In a study of child directed speech to Korean children, Cho (1981), which is cited in Kim (1997), found that of utterances containing transitive verbs with an overt subject and object, on average 5.1% were SVO. If this input is parsed by the learner with O as sister to V, then this input actually favors raising grammars (cf. the analogous discussion on SVO input in Chapter 3). Thus in a more enriched implementation of the model, the effect that adverbial input can have in favoring not raising could be mitigated by this SVO input.

I next consider the second core type of input: input where the verb takes a complement. We have already seen a canonical example of this: SOV input, where the verb takes a direct object. There are other similar types of input we could include, such as the verb taking two internal arguments or the verb taking a complement clause. These kinds of input are also ambiguous for headedness and raising. However, so long as the verb takes an argument as its sister, there is no material difference between these kinds of input and SOV input. Thus these additional kinds of input can push the learner toward a head-final grammar that is either raising or non-raising. What I will focus on instead are verb+auxiliary constructions.

Verb+auxiliary constructions are interesting because they provide a prime example of what might be unambiguous evidence for the parameters here. If such unambiguous evidence were widely available to the learner, it would dramatically change the nature of the learning task. In what follows I will not attempt to exhaustively prove that all constructions involving verbal morphology are ambiguous. Rather, my approach is more to present a methodological blueprint. To the extent that it is successful with some cases, we can then consider applying it to additional examples. I will present an analysis of canonical verb+auxiliary constructions according to which they are structurally more complex than they might at first appear to be. That is, these constructions all involve additional projections. A consequence of having additional structure is that the constructions become ambiguous with respect to head-finality and verb movement for the familiar reasons that we have already seen. This analysis depends on the richness of Korean verbal morphology as a motivating factor for proposing this additional complexity. The approach, then, is to pursue the possibility that all such constructions are ambiguous because they have multiple structural analyses. I will focus on a single canonical type of verb+auxiliary example, but see note 6 for a different type of example relating to unambiguous evidence.

Consider the transitive example in (22a) below. I assume that auxiliaries are a kind of verbal projection, heading a full VP in their own right (cf. Cho 1993 and note 7 below), and that the presence of an auxiliary blocks the main verb from raising to the position of the auxiliary or any higher in the structure. If the auxiliary takes the main verb VP as its complement as per (22b), then the learner would have unambiguous evidence that the main verb VP is head-final: the object is the sister of V, and there is nowhere for V to raise (because the auxiliary blocks raising); therefore VP is head-final.

(22) a. Mary-ka chayk-ul ilk-e po-ass-ta. (Cho 1993: 15)
Mary-nom book-acc read-E try-past-decl
'Mary tried to read a book.'

b.

c.

Verb+auxiliary examples such as (22a) are a robust type of input in Korean. If they can be treated as unambiguous evidence for head-finality, then the nature of the learning challenge this chapter began with changes. The learner could still make use of parameter interaction to learn from the ambiguous input discussed above, but the role of this input would be greatly diminished in the face of the unambiguous input. What I would like to propose is that learning from ambiguous input is vitally important for learners of verb-final language such as Korean because there is reason to think that the structural analysis of (22a) is more complex: auxiliaries embed some larger phrase marker that contains the VP of the main verb, as in (22c). So long as a single projection XP intervenes between the VP of the main verb and the auxiliary, the case of transitive verb+auxiliary input becomes ambiguous and can be thought along the lines of the case of simple SOV transitives. If V raises to X, then XP must be head-final and V could be either final or initial; but if V does not raise to X, then X could be final or initial, but V must be final.

Moreover, the case of intransitive verb+auxiliary input can also be informative in much the same way as SOV input is, with the complement of the auxiliary verb playing a role analogous to the direct object. Suppose the auxiliary is a V that can raise (say, to *v* or T). With either transitives or intransitives, regardless of whether the main verb raises to X (cf. 22c), if the auxiliary does not raise, then the auxiliary must be V-final so as to follow its complement containing the main verb. If the auxiliary does raise to T, then TP must be head-final. The learning scenarios covered here with both main verbs and auxiliaries in both transitive and intransitive clauses are then analogous to what we have seen with SOV input.

What might be the intervening projection XP (or projections) in (22c)? Here I will not try to pin down the exact identity of X, but will simply report on evidence in the literature that point to the existence of functional structure between the main verb and the auxiliary. There are a number of morphemes that can occur between the main verb and the auxiliary. Further, these morphemes are in a selectional relationship with each other. I present two kinds of intervening morphemes here.

First, main verbs are followed by morphemes known as ‘verbal complementizers’ or ‘linking suffixes’, and the identity of these linking morphemes is determined idiosyncratically by the auxiliary (Cho and Sells 1995). For example, the auxiliary ‘try’ requires the linking morpheme *–e* and precludes the linking morpheme *–eya* (23); conversely, the auxiliary verb for ‘must’ requires *–eya* and does not allow *–e* (24). Thus we can say that the auxiliary subcategorizes for the appropriate verbal complementizer, and this relationship can be represented structurally with a syntactic projection; indeed Koopman (2005) assumes these morphemes head their own projections in the syntax.

(23) a. Mary-ka chayk-ul ilk-e po-ass-ta.
Mary-nom book-acc read-E try-past-decl
'Mary tried to read a book.'

b. *Mary-ka chayk-ul ilk-eya po-ass-ta.
Mary-nom book-acc read-EYA try-past-decl
'Mary tried to read a book.'

(Cho 1993: 15)

(24) a. Mary-ka chayk-ul ilk-eya ha-n-ta.
Mary-nom book-acc read-EYA must-pres-decl
'Mary must read the book.'

b. *Mary-ka chayk-ul ilk-e ha-n-ta.
Mary-nom book-acc try-E must-pres-decl
'Mary must read the book.'

(Cho 1993: 15)

Moreover, these linking morphemes subcategorize in turn for the overt expression of additional morphology between them and the main verb. For example, *–eya* allows for the phonologically overt realization of tense morphology on the main verb (25), whereas *–e* does not (26).

(25) a. Mary-ka pap-ul mek-Ø-eya ha-n-ta.
Mary-nom rice-acc eat-pres-EYA must-pres-decl
'Mary must have a meal.'

b. Mary-ka pap-ul mek-ess-eya ha-n-ta
Mary-nom rice-acc eat-past-EYA must-pres-decl

(Cho 1993: 16)

(26) a. Mary-ka chayk-ul ilk-e po-ass-ta.
Mary-nom book-acc read-E try-past-decl
'Mary tried to read a book.'

b. *Mary-ka chayk-ul ilk-ess-e po-ass-ta.
Mary-nom book-acc read-past-E try-past-decl
'Mary tried to read a book.'

(Cho 1993: 15)

One approach to these examples that is consistent with the general syntactic framework being assumed here is to treat (25) on a par with (26) by assuming that there is null tense morphology in (25) (see Sells 1995: 297; cf. the approach in Cho and Sells 1995).[4] The phonological realization of tense morphemes would then be conditioned by their respective selecting verbal complementizers.

[4] Koopman (2005: 618) suggests that *–e* can select for VoiceP, which embeds VP.

I conclude that there is functional structure between the main verb and the auxiliary, although the precise identification of this morphology remains an open question. For the sake of concreteness and as a null hypothesis, let us assume some degree of similarity between the projections dominating the main verb and the auxiliary, where > indicates asymmetric c-command (see also note 9 for further similarities between auxiliaries and main verbs):[5]

(27) T > *v* > Aux/V

In keeping with the null hypothesis, then let us take it for granted that (27) reflects the learner's knowledge of the adult grammar.

The hypothesis in (27) results in a certain uniformity between the structure of the complement of the auxiliary and the structure above of that complement. Thus any time the learner samples a grammar, the same parametric choices will govern these two parts of the structure. It now becomes possible to conflate SOV input with intransitive and transitive verb+auxiliary input for the purpose of this model. Just as raising and non-raising grammatical possibilities cause SOV input to be ambiguous for headedness, the same possibilities make auxiliary verbs and their complements ambiguous in, for all intents and purposes, the same way. And just as the model can settle on certain parameters having head-final values as being the most probable structural analysis for SOV input, the model can also do so for input with auxiliaries (while favoring neither raising nor non-raising) given the structural similarities across the kinds of input. In sum, the exact same set of grammars that are compatible with SOV input is compatible with (in)transitive verb+auxiliary input. Thus these two kinds of input can be treated as the same in the simple 5-parameter grammatical space, and I will not attempt to distinguish verb+auxiliary input from SOV input in the schematic corpus. In the context of the current model, this is not a crucial omission for the learner, and we will see that there is sufficient SOV input for the learner. SOV input will thus be taken as a prototypical type of input that is representative of various constructions in which a verb takes a complement.[6]

---

[5] If the identity of these projections is different, then we might need to expand set of parameters to properly account for verb+auxiliary input. That is, if there is a special projection dedicated to the linking morphemes, we would need to augment the set of functional heads and their respective headedness and raising parameters. However, this should not change the overall learning path outlined in this chapter.

[6] This discussion about auxiliaries is revealing for another construction that has figured prominently in the literature on Korean: so-called conjunctive structures involving *–ko*, which is also one of the verbal complementizers. Examples such as (i) have figured prominently in the literature, with Yoon (1994) claiming that examples such as (ia) instantiate V' coordination (without the second object) or VP coordination (with the second object), while (ib) would be TP coordination.

The discussion about verbal morphology above also relates to claims about the relationship between inflectional morphology and verb movement. I have been assuming throughout that verb movement is not tied to the richness of inflectional morphology. Thus we can note that it is plausible to treat the examples in (25) and (26) as involving multiple inflectional affixes. According to Bobaljik (2000, 2001), such rich inflectional morphology is an indication that the verb must move.[7] In contrast, my claim is that these types of examples are still ambiguous with respect to verb movement. The prediction is that along the lines of the simulation in Section 2 (and, as we will see, the simulation in Section 5), when encountering input such as (25) and (26), the model would sometimes learn a verb raising grammar and sometimes learn a non-verb raising grammar. This claim is supported by the experimental results discussed in Section 4 from Han et al. (2007), who report variability across speakers with respect to verb movement. Let us

---

(i) a. John-i pap-ul mek-ko (kulus-ul) chiu-ess-ta.
John-nom rice-acc eat-KO (dishes-acc) clean-past-decl
'John ate and cleaned the meal/ate the meal and cleaned the dishes.'
b. John-i pap-ul mek-ess-ko chiu-ess-ta.
John-nom rice-acc eat-past-KO clean-past-decl
'John ate and cleaned the meal.'

(Yoon 1994: 252)

The significance of, say, VP coordination for our purposes is that it would provide the learner with unambiguous evidence that VP is head-final. It would also provide evidence against verb-raising: the second verb could not raise out of its conjunct as per Ross's (1967) Coordinate Structure Constraint, and so verbal suffixes must lower and attach onto the verb. However, there is reason to think that as with auxiliaries, the constituents selected by *–ko* embed more syntactic structure. Thus Chung (2005), Lee (2008), and Lee and Tonhauser (2010) observe that the conjuncts in (ii) can have independent temporal interpretations (i.e. the temporal reference of the first conjunct is not dependent on the overt tense morphology that follows the second conjuncts).

(ii) John-i cinan hakki-ey nonmwun-ul ssu-ko, onul machimnay (Lee 2008: 368)
John-nom last semester-at thesis-acc write-KO today finally
colepha-nun-ta.
graduate-pres-decl
'John wrote a thesis last semester, and finally he is graduating today.'

To account for this, Chung (2005) assumes that there is a null T in the first conjunct, making (ia) and (ii) more similar to (ib):

(iii) [[TP] –ko [TP]]

With the more complex structure in (iii), (ia) no longer provides unambiguous evidence and reduces to the familiar case of ambiguity that we have seen with SOV clauses. I have focused here on aspects of the *–ko* construction that are most crucial for the discussion here and have not touched on other pieces of evidence that Yoon (1994) presents in favor of his analysis. For additional discussion against Yoon's arguments I refer the reader to Han et al. (2007) and Chung (2005).

[7] Bobaljik's claim that the verb must move when co-occurring with rich inflectional morphology is couched in a theory of feature-checking that I have not adopted here. I refer the reader to Bobaljik's works for more details on this theory, but I note here that Bobaljik's theory contrasts with my assumptions about the wide availability of operations such as affix lowering or Morphological Merger in the affixation of bound morphemes to the verb.

assume that Korean is indeed a language with rich inflectional morphology, as suggested by (25) and (26). That the modeling and experimental results point toward parametric variability thus provides evidence in support of the general approach pursued here, according to which the richness of inflectional morphology does not provide the learner evidence with respect to verb movement.

I have now considered a representative range of sentence types in Korean, and I have focused on a simple classification of the learner's input into two broad classes of SV and SOV, the former of which is not informative to the learner whereas the latter is. This abstracts away from several more complex factors discussed above, but it presents us with a simplified version of Korean that captures the broad tendencies of the learner's evidence with a sufficient level of detail to see how the model works. This approximation of Korean can of course be made more realistic, but within the scope of the model laid out above, such changes should not affect the overall results reported here.

In (28) I present the frequencies at which the input types are presented to the learner. This repeats what was adopted earlier in (12) as a corpus for the schematic 3-parameter version of the model. The difference is that our understanding of the corpus is now informed by the preceding discussion that has taken into consideration a more diverse range of syntactic constructions.

(28) *Input type and probability*
a. SV $p = .75$
b. SOV $p = .25$

Given that object scrambling is possible in Korean (something which I have not considered in the preceding discussion, though see Section 4 for some discussion of movement of the object), it is likely that (28) overestimates the frequency that the object occurs linearly adjacent to the verb. Recall, though, that for the syntactic parameters at hand, I am assuming that the complements of auxiliaries are informative to the learner in much the same way that OV sequences are. Were verb+auxiliary sequences to be included in (28), that would be comparable to increasing the proportion of SOV input. What we see in the results that follow is that even without augmenting SOV input with verb+auxiliary clauses, 25% is a sufficiently large proportion of the input for the model to converge on either raising or non-raising grammars that have head-final values for the projections under consideration.

3.3 Predictions for the model

Given the discussion above, expected results from running the model are presented more formally below. These predictions follow the same contours as those for the 3-parameter model in (13) and (14). Again, they are formulated to look at averaged results from running the model multiple times. Multiple runs of the model can be used to simulate different speakers in a population learning the language. The predictions are

based on the weight a given parameter's values have relative to each other across multiple runs. A high proportion of the weight for a value can be taken to indicate that the model has learned that value, while a low proportion indicates the model has not learned the corresponding value. (29a) predicts a point in the learner's development at which the model adopts a consistently head-final grammar, while (29b) predicts that at that same developmental point there will be variability as to whether this grammar is raising or non-raising. (29a) entails that on average in any particular run the weights for the headedness parameters will approach 100% for head-final (30a).

(29) *Predicted weights of parameter settings (averages for multiple runs)*
a. Weights near 100% for {V-fin, *v*-fin, T-fin}
b. Weights not near either 100% or 0% { [+V-*v*], [–V-*v*], [+*v*-T], [–*v*-T]}

(30) *Predicted weight of parameter settings for a single run*
a. Weights near 100% for {V-fin, *v*-fin, T-fin}
b. Weights near 100% for {[+V-*v*] or [–V-*v*]} and {[+*v*-T] or [–*v*-T]}

Additionally, for any particular run of the model we predict (30b) for raising to *v*. Although we predict variability across runs or speakers in (29b), with (30b) we predict a point in the learner's development at which this variability is stable within a single run/speaker. As in Section 2.2, we can note that the variability predicted here runs counter to the null hypothesis, according to which there is no grammatical variability.

I postpone reporting on results of running the model in the 5-parameter hypothesis space until Section 5. In the following section I compare the analysis of variability in the learning model with the syntactic analysis of Korean in Han et al. (2007). The experimental results of Han et al. can be taken to support the general approach to variability pursued here (and thus as additional support against the null hypothesis of no variability), but their analysis of Korean is seemingly at odds with the conclusions drawn in this section. I propose a modification of their analysis that attempts to unify their approach with the model of Korean discussed here. With this modification in mind, results of running the model are given in Section 5, along with more technical discussion of the priors and update procedure.

4. Han et al. (2007) and the current model: A deeper look at modeling stable variability

In the preceding sections we have seen how the ambiguity in Korean can lead to grammar variability in the framework of this model. In this section I discuss the experimental and acquisition work in recent papers such as Han et al. (2007). The purpose of this discussion is twofold. First, Han et al.'s experimental results provide empirical support for the variability that we expect to see when we run the learning model. I review the relevant results from this literature in Section 4.1. Second, a comparison of the model's parameter space with Han et al.'s analysis raises several points of difference that could be problematic for modeling variability. These are discussed in Section 4.2. Although both the experimental work and the learning model dovetail on the point of variability, there is a crucial difference regarding the position of the object in the analyses of Korean underlying them. To the extent that these different approaches are attempting to capture the same phenomenon, enriching the scope of the model to include aspects of Han et al.'s analysis is thus a clear next step for the proof-of-concept model. However, if we were to apply Han et al.'s analysis wholesale to the learning model, this difference would lead us to expect that there would not be variability across speakers. Such a difference would pose an immediate problem in trying to model variability and leads to the practical question of whether the proof-of-concept model is viable given an obvious extension to it.

In Section 4.3 I offer some speculative remarks on how both analyses can be slightly modified so as to bring them more closely in line with each other. These modifications will allow us to keep the model of Korean from Section 3 essentially in its current format. Results from running the model that illustrate grammatical variability are then given in Section 5.1. It thus is still an open question as to whether the model here accurately models the variability seen in Korean. At the very least the current proposal is an attempt to model Korean variability; Han et al. provide no formal model of how this variability occurs in the population of learners. Further, the proposal illustrates the kind of conditions necessary to converge on stable variability with a probabilistic learner operating with multiple interacting parameters and learning from entirely ambiguous input. To my knowledge this has not been demonstrated before and thus constitutes an important first step in modeling empirical phenomena such as the variability in Korean. This is discussed after the results in Section 5.2.

4.1 Review of Han et al. (2007)

Han et al. (2007), which I will refer to as HLM, propose a novel empirical test for detecting whether there is verb movement in Korean. This involves looking at the scope of so-called 'short form negation' with respect to a quantified object. Short form negation immediately precedes the verbal complex (31a) and contrasts with 'long form negation',

which occurs post-verbally and requires the presence of the auxiliary *ha–* 'do' (31b). HLM obtained similar results for long negation, but for expository purposes, I will focus on short negation. I refer the reader to HLM for more on long negation, as well as note 9 for a short discussion on long negation.

(31) a. Toli-ka an tten-ss-ta. (HLM: 14)
Toli-nom Neg. leave-past-decl
'Toli didn't leave.'

b. Toli-ka tten-ci ani ha-yess-ta. (HLM: 13)
Toli-nom leave-CI Neg. do-past-decl
'Toli didn't leave.'

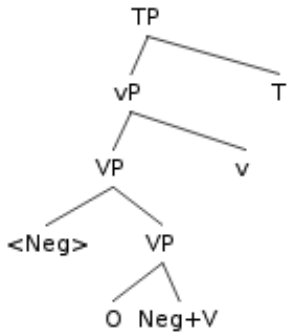
According to HLM's analysis, the scope of short negation is tied to the position of the verb: if the verb raises, then negation will have high scope; if the verb stays in-situ, then the verb will take low scope. The difference in scope can be detected with a quantified object, which is assumed to move to a fixed position. The landing site of verb movement is above this object position, and so the verb scopes over the object if it moves. The base position of the verb is structurally lower than this object, and so the object takes wide scope over negation if the verb does not move. HLM then conduct an experiment with adults and children as participants, using the truth-value judgment task (Crain and Thornton 1998) to determine native speaker judgments of the scope of negation. Their findings and those in Han et al. (to appear) show that there is a high degree of stable variability in the judgments of speakers concerning the scope of negation. They conclude that this reflects a more fundamental variability across speakers: whether they have a adopted a raising or a non-raising grammar.

According to HLM, an important point concerning this variability is that it is due to the ambiguity of the learner's primary linguistic data. As has been discussed in the preceding sections, the input to a Korean learner is ambiguous as to whether there is verb raising. In principle, then, the scope of negation could vary across speakers (but remain constant within speakers) as a function of this ambiguity with respect to verb raising. But might the scope of negation be an important type of unambiguous evidence to the learner, thereby resolving the ambiguity as to whether there is verb raising? I follow HLM in assuming that evidence containing negation, a quantified object, and an appropriate context that would clearly distinguish the position of negation (i.e. Context 1 for (37) below) is sufficiently rare in the input to the learner that it has little to no role in the average learner's development. How rare is this evidence? HLM do not conduct a corpus analysis, and it would be desirable to do so. However, their assumption is a reasonable one given that most speakers do not exhibit within speaker variability. Consider the alternative, which would involve parameter setting from unambiguous evidence containing negation. According to this idea, there would be two contrasting kinds of unambiguous input, and Korean learners would set their verb movement parameter(s)

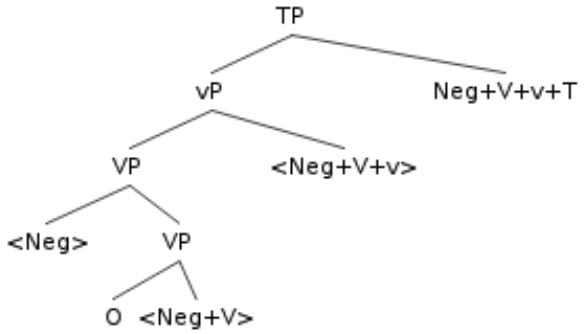
based on which kind of unambiguous input they were exposed to. Let us also assume that this evidence is available to all speakers. If this were the case, the fact that most speakers do not show within speaker variability is puzzling. If both types of unambiguous evidence were robust enough to allow for parameter setting, why is that most speakers consistently have grammars with one parameter value or the other, but not some probabilistic mix of both? After all, by hypothesis they would have been on average exposed to a sufficient amount of both kinds of unambiguous evidence. As discussed in Section 5.2, the stable variability of Korean speakers contrasts with diachronic cases of within speaker variability, which can (at least sometimes) be characterized by robust frequencies of contrasting types of unambiguous evidence. Given these considerations, I will assume that this kind of unambiguous evidence is insufficiently attested in the learner's input to set a parameter. Thus I maintain the claim that for all intents and purposes, all Korean input is ambiguous for the parameters I look at in this chapter.

There are three core components of HLM's analysis of the scope of short negation that ties it to the structural position of the verb. First, HLM assume that short negation is a clitic that is first merged into the structure by adjoining to VP before attaching to the verb in the syntax. The syntactic position of negation is thus tied to the position of its host: if the verb does not raise, then negation will remain in VP (32a); if the verb raises, say to T, then short negation will be in T (32b):

(32) a.



b.



Second, HLM claim that objects undergo obligatory raising from their initial position in the VP. For concreteness, let us assume that objects are merged into the

structure as the sister of V. The primary evidence that objects cannot remain in such a position comes from examples such as (33a). There is a small class of modifiers (see Lee 1993) that must appear left-adjacent to the verb. One of these modifiers is *cal* 'well'; this contrasts with a modifier such as *pelsse* 'already', which has no such restriction (33b).

(33) a. (*cal) Chelswu-nun (*cal) sayngsenhwoi-lul (cal) mek-nun-ta. (Lee 1993: 434)
(well) Chelswu-nom (well) raw fish-acc (well) eat-pres-decl
'Chelwsu eats raw fish well.'

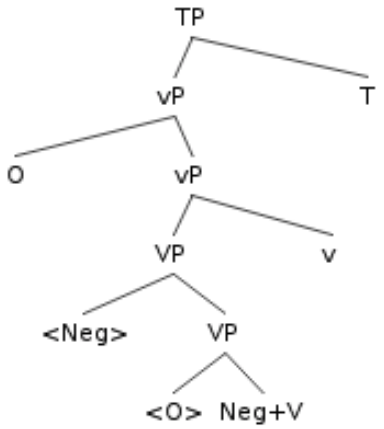
b. (pelsse) John-un (pelsse) yenge kongpwu-lul (Hagstrom 2002: 219)
(already) John-nom (already) English studies-acc
(pelsse) machi-ess-ta.
(already) finish-past-decl
'John has already finished his English studies.'

If the object's base position is the sister of V, then if it could remain in its base position in (33a), we would expect it to be able to be left-adjacent to the verb, contrary to (33a). HLM thus claim that the object must raise to a higher position; they assume this is the specifier position of some higher functional projection, which they label FP. Nothing in their analysis hinges on the identity of F; to make their proposal more directly comparable to the learning model, I will label this projection *v*P. Object raising is illustrated schematically below.

(34) *Object Raising*
[$_{TP}$ [$_{vP}$ O [$_{vP}$ [$_{VP}$ ~~O~~ V] *v*] T]

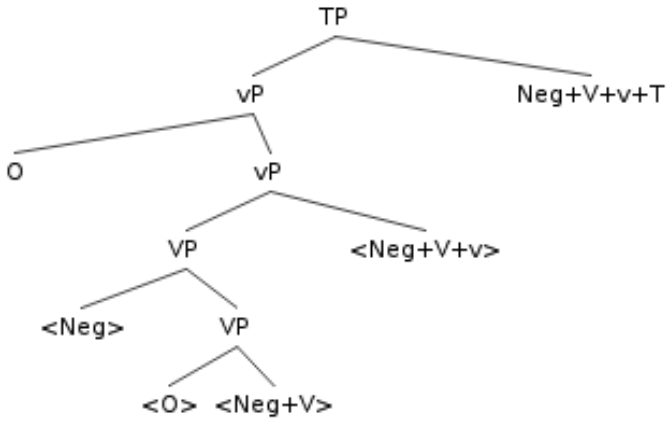
Third, HLM adopt a widely held assumption in the literature on Korean and Japanese (Joo 1989; cf. Kuroda 1970) that semantic scope at Logical Form in Korean is determined by the c-command relations of scope-bearing elements in overt syntax. The implication of this, combined with (32) and (34) is as follows. If the object raises (34), but the verb does not (32a), then the object will have wide-scope over negation:

(35) *Object > Negation*



If the object raises, and the verb raises as well (32b), then negation will have wide-scope over negation:

(36) *Negation > Object*



Under this analysis, judgments about the scope of negation provide a clear window onto the syntax of verb raising in Korean.

HLM use this conclusion as a diagnostic for verb raising in a truth-value judgment task that tests the scope of negation with respect to a universally quantified object. Participants are presented with a context (acted out with toys) and are then presented with a sentence such as (37). The participant is then asked whether (37) is true (a 'yes' response) or false (a 'no' response) given the context.

(37) Khwukhi Monste-ka motun khwukhi-lul an mek-ess-ta. (HLM: 28)
Cookie Monter-nom every cookie-acc Neg eat-past-decl
'Cookie Monster didn't eat every cookie.'

There are two contexts used to test (37). To test the neg > ∀ reading, Cookie Monster eats two of three cookies that he is given; let us call this Context 1. To test the ∀ > neg reading, Cookie Monster eats none of the cookies; let us call this Context 2. Note that both scope readings are possible given Context 2. A speaker that has verb raising will accept (37) for both contexts. In contrast, a speaker that has no verb raising will only accept (37) given Context 2 and will reject Context 1. What crucially distinguishes a raising from non-raising grammar, then, is the rejection of Context 1, i.e. rejecting neg > ∀. Under neither grammar would a speaker reject ∀ > neg.

Predictions are as follows. If all speakers have verb raising, then we expect a very low rejection rate of neg > ∀. If no speakers have verb raising, then we expect a very high rejection rate of neg > ∀. Finally, if there is variability across speakers, we expect a large group of speakers to accept neg > ∀, and the remaining speakers reject it.

20 adults and 15 children were tested on sentences like (37). Results are given below in Table 4.2 (which is based on HLM's Table 6; HLM: 30). Table 4.2 indeed

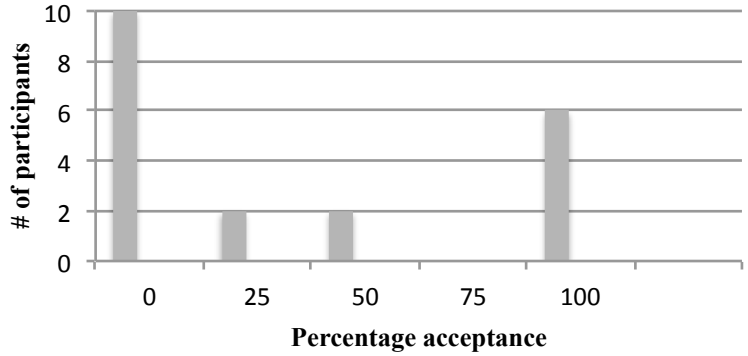
shows that there is variability in adults. The adults tested accepted neg > ∀ (indicating a verb raising grammar) only around one-third of the time. As expected, there was essentially no difference in acceptance for ∀ > neg.

Table 4.2 Mean percentage of acceptances for adults

Scope	**Acceptance rate**
neg > ∀ object	37%
∀ object > neg	98%

Further, this variability is highly consistent within an individual speaker's grammar. As shown in Figure 4.1 (which is based on HLM's Figure 4; HLM: 31), speakers largely split into two groups: one group that consistently rejects Context 1, and one group that consistently accepts it. Thus speakers are not typically vacillating back and forth between accepting and rejecting the context; their grammars are stable.[8]

Figure 4.1 Number of participants accepting neg > ∀ object *for adults*



We see similar results for the children that were tested. These children were all between the ages 4;0 and 4;11 (mean 4;5). Context 1 is accepted only around one-third of the time, as shown in Table 4.3 (which is based on HLM's Table 8; HLM: 36). The children's grammars are also stable, with the population of participants falling into two groups. 9 of the children consistently rejected Context 1, while 5 children always accepted it. Only one child accepted it 50% of the time.

[8] It is not clear what can be said about those participants who accepted the context 25%, 50%, or 75% of the time. Their acceptance-rate could reflect a small sample size of sentences in the TVJT. With a larger sample size, it is possible that their acceptance-rates would approach 0% or 100%. In the current model, a 25/50/75% acceptance-rate could be taken to represent a conservative learner for which it is difficult for probabilistic tendencies in the parameter space to push one way or the other. Above all, it is to be noted that these participants represent a relatively small proportion of the population that was tested.

Table 4.3 Mean percentage of acceptances for children

Scope	**Acceptance rate**
neg > ∀ object	36.67%
∀ object > neg	81.67%

These results can be taken to show that the population of Korean speakers is split between a group that has learned to raise the verb and one that has not. Moreover, this variability is stable over time. In a subsequent experiment, Han et al. (to appear) tested the same 26 adults two times, the second time one month after the first, on a similar truth-value judgment task. This was done to investigate whether speakers are using a single grammar over time, or whether they randomly adopt a value for verb raising at the beginning of each experimental session and use it consistently throughout the session. If speakers are oscillating between grammars, then we expect to see differences when comparing a given speaker's responses from different experimental sessions. Han et al. (to appear) report that when re-testing the speakers, no evidence could be found of the participants having changed their responses. This second experiment further supports the claim that participants fall into discrete sub-populations that have parametrically different grammars.

4.2 Comparison of Han et al. (2007) and the current model

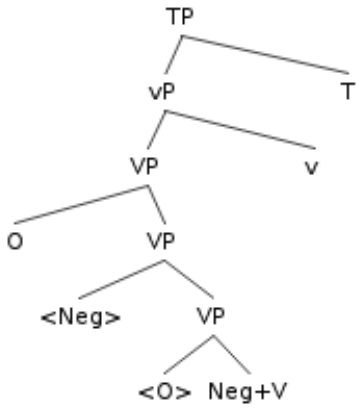
The experimental results of Han et al. (2007) and the results of the model here both point in the same direction. Given data ambiguous for verb raising, some proportion of the population will learn a raising grammar, while the rest will learn a non-raising grammar. Further, the analysis of HLM assumes largely similar syntactic parameters to those of the model run here. However, there is a fundamental difference in the syntax of the two. Han et al. assume that the object raises to Spec*v*P (what they call SpecFP), as in (35)/(36), whereas in the learning model here, I have assumed the learner treats the object as remaining the sister to the verb. The difference between having object raising or not is crucial. With object raising, there is no longer symmetry between verb-raising and non-verb-raising grammars; non-verb-raising grammars will be favored. I illustrate the significance of this difference in this section.

Before discussing the significance of object raising, I address another related difference, but one that I assume is not crucial. For HLM the locus of verb raising is T; whether or not the verb raises to T with negation determines the scope of negation. In contrast, in Section 3.2 I raised the possibility that we can distinguish learners of raising versus non-raising grammars on the basis of whether the verb raises to *v* (cf. the discussion of (20)). As shown in (19), raising to T is less likely than raising to *v* because

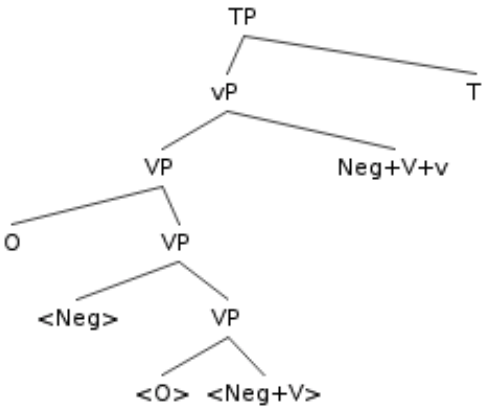
raising to T is dependent on learning that there is a step of shorter movement to *v*. So as to maximize the likelihood that the model can learn verb raising, I have focused on whether it learns any verb movement at all, i.e. whether it learns there is verb movement to *v*. Yet at first pass, it might appear that by only raising to *v*, the verb is not raising high enough for the appropriate scope of negation. Crucially for Han et al., the verb can asymmetrically c-command the object when raising to T, and the object can asymmetrically c-command the verb when the verb remains in VP. When the verb raises to *v*, this is not so for the following reasons. First, if we assume that the object raises to Spec*v*P and the verb to *v*, then neither the object nor the verb will asymmetrically c-command the other. It is no longer clear, then, whether the verb+negation can systematically outscope the object for speakers who have verb raising. Second, if we assume that the object does not raise and remains the sister to V, a similar complication ensues. When the verb remains in-situ, neither the object nor the verb will asymmetrically c-command the other.

This apparent impasse can be remedied by a simple modification to HLM's analysis. All that is crucial is that the object raise to some position that is not sister to V. Instead of raising to Spec*v*P/FP, it is sufficient for the object to adjoin to VP. Recall that the motivation for object raising was to account for the word order facts with adverbials in (33); adjunction to VP will capture these facts. Maintaining some kind of object raising also correlates with some additional acquisition data discussed in Section 4.3, where it appears that some children have difficulty learning object raising. With the object adjoining to VP, then if the verb stays in-situ, the object will asymmetrically c-command it (38a), and if the verb raises to *v*, it will asymmetrically c-command the object (38b).

(38) a.



b.



I note that this modification is readily compatible with HLM's analysis. First, HLM give no specific reason for the object to move to Spec*v*P/FP *per se*. This is consonant with their abstracting away from the identity of the projection, simply calling it FP. Second, the role of this projection is largely twofold: (a) to allow for the site of verb raising to asymmetrically c-command the object, while the object can asymmetrically c-c-command the base position of the verb; and (b) to allow for low adverbials to appear between the object and the verb. This role can be played effectively by VP-adjunction of the object instead. As shown in (38), the same asymmetric c-command relations are preserved. Further, low adverbials can now adjoin below the VP-adjoined position of the object. Along the lines of HLM, then, I will assume that in the adult grammar in Korean, the object raises and adjoins to VP.[9]

---

[9] HLM obtained similar results regarding the scope of negation and a quantified object with long negation. In long negation, which was illustrated in (31b), the negation marker is pronounced after the main verb and must co-occur with the auxiliary *ha-* 'do'. Accounting for the results with long negation would involve a similar proposal to what has been discussed in the text. First let us consider the structural position of long negation and *ha-* 'do'. HLM assume that long negation *ani* adjoins to *ha-*, which heads a projection above the main verb, but below TP, as in (i), where > indicates c-command:

(i) T > Neg+*ha-* > V

However, Sells (1995) treats *anh-*, which is the contracted form of *ani*+*ha-*, as an independent auxiliary. If we follow Sells in assuming that the exponence of long negation is in fact an auxliary, then we automatically account for *ha-* always co-occurring with long negation in the same syntactic position. As for the base position of long negation, there is reason to think that it occurs in the same base position as other auxiliaries. Thus, Sells (p. 305) notes that as with other auxiliaries, honorifcation marking can co-occur on the main verb and *anh-*. Recall from (27) that I assume that auxiliaries are generated in a position below TP. This would mean that the auxiliary in long negation is not a 'dummy' that is inserted in T to support tense; rather, it appears below T like other auxiliaries. Support for this comes from examples where we see double long negation, as in (ii).

(ii) Chelswu-nun chayk-ul ilk-ci ani ha-ci ani ha-ess-ta. (Hagstrom 1995: 18)
Chelswu-top book-acc read-CI Neg.-do-CI Neg.-do-pst-decl
'Chelswu didn't not read the book.' (= Chelswu read the book.)

The fact that there are two *ha-* morphemes is unaccounted for if *ha-* were inserted only to support overt tense morphology. If long negation necessarily involves an auxiliary, though, the construction in (ii) is

Returning to the significance of object raising, if the parameter space were increased to 6 binary parameters with the inclusion of a parameter for object raising, then there would be 64 logically possible grammars. 40 of these grammars are compatible with SOV input. These are given below according to whether or not the learner chooses a grammar with object raising. If the learner chooses no object raising (39a), then there are 16 compatible grammars. These grammars have the same values for the remaining 5 parameters as above in (18), for which object raising was not considered. If the learner chooses a grammar with object raising (39b), the same grammars as in (39a) are compatible, as are an additional 8 grammars. These additional grammars are in boldface below. The 8 additional grammars are all [–V-*v*, V-init] grammars. This is because if the verb stays in-situ and the object now adjoins to VP, then even if the VP is head-initial, the verb will still appear string-finally.

---

again accounted for automatically. I assume then, that long negation is generated in a position, say AuxP, above the VP of the main verb, but below TP.

Variability in scope facts then proceeds as follows. First, consider the quantified object scoping over long negation. Parallel to the case of short negation, by hypothesis the long negation auxiliary has not raised from AuxP. The object can scope over in the in-situ long negation if it raises and adjoins to AuxP. If the object has adjoined to AuxP, then long negation can scope over the object if it has raised to a higher projection such as *v*P. The fact that HLM found similar judgments within speakers for both types of negation suggests that the same parameter regulates verb raising and auxiliary raising. For example, with a value of [+V-to-*v*] a speaker would raise both verbs and auxiliaries to different projections of *v*. This would be another instance of parallelism between main verbs and auxiliaries.

This account of the variable scope judgments with long negation is highly parallel to that of short negation, with the following difference. In the case of short negation, I proposed that the object adjoins to the VP of the main verb. Now with long negation, I have proposed that the object adjoins to the AuxP above the main verb. This suggests that the precise formulation of the parameter for object raising allows the object to target not just VP but also AuxP for adjunction. This is not obviously problematic given the discussion both here and in the main text that verbs and auxiliaries pattern very similarly. The hypothesis, then, is that object raising would adjoin to the projection of the highest among these two kinds of heads. Whether this is indeed the correct characterization of object raising remains a question for future research.

(39) *Grammars compatible with SOV input (6-parameter hypothesis space)*

a. <u>*No object raising*</u>	b. <u>*Object raising*</u>
1. [+V-*v*, +*v*-T, T-fin, *v*-in, V-in]	1. [+V-*v*, +*v*-T, T-fin, *v*-in, V-in]
2. [+V-*v*, +*v*-T, T-fin, *v*-in, V-fin]	2. [+V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
3. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-in]	3. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-in]
4. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]	4. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]
5. [+V-*v*, –*v*-T, T-in, *v*-fin, V-in]	5. [+V-*v*, –*v*-T, T-in, *v*-fin, V-in]
6. [+V-*v*, –*v*-T, T-in, *v*-fin, V-fin]	6. [+V-*v*, –*v*-T, T-in, *v*-fin, V-fin]
7. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-in]	7. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-in]
8. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]	8. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]
9. [–V-*v*, +*v*-T, T-in, *v*-in, V-fin]	9. [–V-*v*, +*v*-T, T-in, *v*-in, V-fin]
	10. [–V-*v*, +*v*-T, T-in, *v*-in, V-in]
10. [–V-*v*, +*v*-T, T-in, *v*-fin, V-fin]	11. [–V-*v*, +*v*-T, T-in, *v*-fin, V-fin]
	12. [–V-*v*, +*v*-T, T-in, *v*-fin, V-in]
11. [–V-*v*, +*v*-T, T-fin, *v*-in, V-fin]	13. [–V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
	14. [–V-*v*, +*v*-T, T-fin, *v*-in, V-in]
12. [–V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]	15. [–V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]
	16. [–V-*v*, +*v*-T, T-fin, *v*-fin, V-in]
13. [–V-*v*, –*v*-T, T-in, *v*-in, V-fin]	17. [–V-*v*, –*v*-T, T-in, *v*-in, V-fin]
	18. [–V-*v*, –*v*-T, T-in, *v*-in, V-in]
14. [–V-*v*, –*v*-T, T-in, *v*-fin, V-fin]	19. [–V-*v*, –*v*-T, T-in, *v*-fin, V-fin]
	20. [–V-*v*, –*v*-T, T-in, *v*-fin, V-in]
15. [–V-*v*, –*v*-T, T-fin, *v*-in, V-fin]	21. [–V-*v*, –*v*-T, T-fin, *v*-in, V-fin]
	22. [–V-*v*, –*v*-T, T-fin, *v*-in, V-in]
16. [–V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]	23. [–V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]
	24. [–V-*v*, –*v*-T, T-fin, *v*-fin, V-in]

(40) *Partial summary of grammars in (39)*

a. –V-*v* grammars: 60%	b. V-in grammars: 40%
+V-*v* grammars: 40%	V-fin grammars: 60%

The effect of these additional 8 grammars is two-fold. They now shift the learner strongly toward a non-raising grammar. With a favorable majority of compatible grammars being [–V-to-*v*] (40a), symmetry between raising and non-raising grammars is lost, and we no longer expect variability with respect to verb raising. Second, the additional 8 grammars are all V-initial. They slightly decrease the majority of V-final grammars from two-thirds in (21) to 60% (40b). Thus we would still expect the model to learn a V-final grammar, but the model is now expected to learn that Korean does not have raising.

In sum, object raising radically changes the predicted learning outcomes for the model. Given such a significant difference between the analysis of Han et al. and the results of the learning model, a natural question is whether the results of the learning model accurately reflect the experimental results of Han et al. In the following section I draw on additional acquisition data from Korean children to propose a possible way of

unifying the model's results and those of Han et al.

4.3 Toward a unification of Han et al. (2007) and the current model

In Section 4.2 I discussed how the crucial syntactic difference between HLM's analysis and that of the model concerned the position of the object. In HLM's analysis, the object raises out of the VP. In the modification I proposed for that analysis, the object raises and adjoins to VP. Importantly, under these analyses the object does not remain in-situ as the sister of the base position of V, as it does in the learning model above. I also discussed in Section 4.2 how the inclusion of a parameter for object raising now pushes the learner to favor a non-verb raising grammar. There is thus a tension between object raising and verb raising. Is it possible to have a parameter for object raising while still learning a grammar with verb raising? In this section I look at additional acquisition data from Korean children and make a proposal concerning the structural position of objects during children's development. I propose that there is a delay in children's learning object raising such that children can first learn whether there is verb movement along the lines I have already described, after which they learn that there is object raising. The effect of this proposal is thus to bring the structural analysis of objects in the learning model in line with the analysis in Han et al. (2007). The discussion in this section is speculative in nature, but the acquisition data is highly suggestive. Whether or not the learning model can be fully unified with Han et al. awaits further research and testing of the model.

To begin, assuming that adults do raise the object, let us assume that there is a syntactic parameter governing this raising. Accordingly, Korean adults have set this parameter positively so that there is object raising. But what about children learning Korean? Have children learned the target setting for this parameter? And to what extent does this parameter play a role earlier in development when the children are learning other parameter settings?

To investigate these questions, let us consider child utterances that contain both a direct object and short form negation *an*, which must appear immediately before the verb. Recall from Section 4.1 that Han et al. assume short form negation adjoins to the VP edge before cliticizing to the verb in the syntax. Before any object raising or cliticization, the structure of the VP would be (41a). Assuming that object raising targets the leftmost edge of the VP, then regardless of whether there is cliticization (41c) or not (41b), object raising will place the object to the left of short form negation.

(41) a. [$_{VP}$ Neg [$_{VP}$ O V]]
b. [O [$_{VP}$ Neg [$_{VP}$ ~~O~~ V]]] *Object Raising*
c. [O [$_{VP}$ ~~Neg~~ [$_{VP}$ ~~O~~ Neg+V]]] *Object Raising + Neg Cliticization*

According to this analysis, if the child has learned object raising, then in the relevant child productions we expect the object to precede short form negation. If the child has not set the parameter governing object raising, it should be possible for the child productions

to have short form negation incorrectly preceding the object. Turning to the acquisition data, various researchers (e.g. Cho and Hong 1988 and Kim 1997) have indeed observed non-target spontaneous child productions (spanning ages 1;11 to 3;5) in which short form negation precedes the direct object. Examples are given below. Non-target child productions are indicated with #.

(42) a. #…an pyeng kkay-ss-e. (J: 2;3)
…Neg bottle break-past-decl
'(I) did not break the (milk) bottle.'
(cf. …*pyeng an kkay-ss-e*.) (Kim 1997:378)

b. # An kkwum kkwu-ese… (P: 2;6)
Neg dream dream-because…
'Because (I) did not dream…'
(cf. *Kkwum an kkwu-ese*…) (Kim 1997: 377)

This acquisition evidence bears directly on the position of the object. In productions such as those in (42), I assume that the object remains in-situ as sister to V. Thus there is evidence suggesting that there is a developmental stage during which children have not yet correctly set the parameter governing object raising.

One way of accounting for these productions involves the following hypothesis. Initially, children never raise the object. Subsequently there is a stage during which children gradually learn that the object raises. Kim (1997) reports that the non-target productions co-occur with target productions, indicating gradual acquisition of the target parameter setting. At present, I have no principled explanation for why children initially would never raise the object. However, such a proposal not only accounts for the production errors, it can serve as a bridge between the results of the learning model and Han et al. Consider the following developmental path. At an early developmental stage the child never raises the object and always treats it as the sister of V. During this stage, the child learns head-finality and whether or not there is verb raising as per the discussion in this chapter. At a later age, the child then correctly learns there is object raising and sets the relevant parameter accordingly. As the headedness and verb-raising parameters have already been set, object raising will have no effect on those parameters. Thus when the child is later tested on the scope of negation, he or she can make use of the asymmetric c-command relations that hold between the displaced object and the raised/in-situ verb. This proposal is also in line the developmental fact that the children who produce errors of the sort in (42) are in general younger than the children in HLM's study, the youngest of which was 4;0.[10]

[10] Interestingly, Anderssen et al. (2011) observe a delay in Norwegian children learning object shift. This was observed in both spontaneous productions and targeted elicitations. Josefsson (1996/2013) found similar results for a delay in learning object shift with Swedish children. Given the rough similarity between object shift in Scandinavian languages and object raising in Korean (both involve highly local

One way of modeling such a developmental course is to assume that the parameter governing object raising initially has a strong prior against object raising. Up to this point, I have adopted the null hypothesis in assuming that competing parameter values have equal priors. With a strong prior against object raising, the hypothesis is that early in development, the learner would treat the object as essentially always being in-situ as sister to V and could thus learn whether or not there is verb raising. Only later in development after a sufficient amount of input indicating that the object had raised would the learner be able to overcome the prior and gradually set the parameter correctly. A strong prior against object raising amounts to something like a default value for this particular parameter. Wherever possible, I have attempted to account for children's development without the use of biased priors, but here no account of the errors in (42) with equal priors is forthcoming. As such, the use of a strong prior against object raising is a complication to the theory, and one that should be explained, but it does find some motivation from the acquisition facts.

In this section I have outlined a proposal for how an object raising parameter could be included in the model while still being able to capture variability in verb raising across learners. This approach is admittedly rather speculative, and it remains to be seen whether it could be successfully modeled. Nevertheless, the hope is that it paves the way toward a unification of Han et al., the learning model, and the non-target child Korean productions.

5. Results and discussion

In Section 5.1 I report results of running the learning model in the 5-parameter space and compare these results with the findings in Han et al (2007). I use the same priors and update procedure as in the 3-parameter version of the model; a description of this can be found in Section 2.2. In Section 5.2, I then compare the current case with other examples of syntactic variability that have been modeled in the literature. The case of Korean is distinctive in that we see relatively stable variability resulting from an entirely ambiguous corpus.

5.1 Results of the 5-parameter model

Results of running the model 15 times are given below in Table 4.4. These results show the average weight for opposing parameter values after running the model for an average of approximately 1,650 tokens of input. These averages are based on what proportion a certain value's pseudo-count total is out of 10 (which is the sum of the pseudo-counts of both of a given parameter's values). A high average represents a point

---

movement of the object), the finding of a cross-linguistic delay in acquiring similar phenomena is suggestive that the approach being taken in the text is on the right track.

in the learner's development at which the model has learned the corresponding parameter value.

The results in Table 4.4 with 5 parameters are highly similar to those reported for the 3-parameter model (cf. Table 4.1 in Section 2.2). Thus we see high averages near 100% across all runs of the model for V-final, *v*-final, and T-final. Depending on the run of the model, though, we have different results for the raising parameters. On some runs, one or both of the raising parameters will have a high average, while on other runs one or both will have a low average (the shaded cells in Table 4.4). These results confirm the predictions in (29) and (30). While the model is successful at consistently learning head-final values for VP, *v*P, and TP, it varies on whether it learns a raising grammar or a non-raising grammar, with a non-zero probability of learning either raising or non-raising. Recall that for the purpose of modeling variability here, it is sufficient that some proportion of the time the model learns that there is *some* raising, even if it is only to *v*. In only one run of the model is a [V-*v*, +*v*-T] grammar learned. Indeed, as discussed in Section 3.1, a grammar that raises the verb to T is only expected to be learned a subset of the time that verb movement is learned. In Section 4.2 I discussed how simply raising to *v* is compatible with Han et al.'s (2007) analysis and experimental results concerning variability in Korean.

Table 4.4 Average proportions of probabilistic weights for parameter values (Average of approximately 1,650 tokens of input per run)

Parameter Value	**Proportion**	**# Runs**
[V-fin]	.9921	15
[V-init]	.0079	
[*v*-fin]	.9338	
[*v*-init]	.0662	
[T-fin]	.9476	
[T-init]	.0524	
[–V-*v*]	.9232	13
[+V-*v*]	.9427	2
[–*v*-T]	.9459	6
[+*v*-T]	.8686	9

And similar to the 3-parameter model, in the 5-parameter model the learner is never pushed strongly toward a head-initial value for any head-complement parameter throughout the course of learning. In general, the weights for head-initial parameter values go steadily down, and the weight for a head-initial value never rises above or comes close to 75%. During two runs of the model the weight for [–*v*-T] swings from above 75% to below 25%, though at the time of high initial weights for [–*v*-T] during these runs, the model has not yet assigned high weights for all the head-complement ordering parameters. There is thus almost no variability in parameter setting throughout

the course of learning, and none for parameters concerning head-complement order. Again the model is consistently learning head-final values for certain projections, but the grammar it learns either has or doesn't have verb raising. And thus we again see evidence, coupled with the experimental results of Han et al. (2007) reviewed below, against the null hypothesis that there is no grammatical variability across speakers.

One question raised by these results is why a verb raising grammar appear to be less favored than a non-verb raising grammar. When we look more carefully at how the parameters interact, we see that there is an additional, subtler effect of parameter interaction, which actually favors learning a non-raising grammar. To see this, consider again the generalizations, repeated in (19) below, regarding the grammars in (18), which are compatible with SOV input in the 5-parameter hypothesis space.

(19) *Summary of grammars in (18)*

a. V-fin grammars: 75%
   V-in grammars: 25%

b. *v*-fin grammars: .625%
   v-in grammars: .375%

c. T-fin grammars: .625%
   T-in grammars: .375%

d. +V-*v* grammars: 50%
   –V-*v* grammars: 50%

e. +*v*-T grammars: 50%
   –*v*-T grammars: 50%

f. [+V-*v*, +*v*-T] grammars: 25%

Among all parameter values, the learner is pushed most strongly toward learning V-final. The greatest majority of grammars containing any particular parameter value contain V-final. Developmentally, this means that early in the learning process, the learner is more likely to reinforce V-final than any other parameter value. Roughly, a V-final grammar is learned first. However, when we consider the 12 grammars that are V-final, we see that the majority of them (8 of 12) are actually [–V-*v*]. Further, all compatible [–V-*v*] grammars are V-final. Thus as the learner is pushed toward V-final, as is predicted, there will be a slight edge in favor of a non-raising grammar.

We can characterize the above scenario of favoring a non-raising grammar as resulting from what I will call *secondary parameter interaction*: given that the learner adopts parameter value $P_1(x)$, some other value $P_2(y)$ occurs in more compatible grammars than $P_2(x)$. In the 5-parameter model we see that given that V-final is learned before any other value, a non-raising grammar is more likely than a raising grammar. Overall, a non-raising grammar is more likely because a V-final value is more likely to be assigned a high weight than any other parameter value. In this way, learning one parameter value can act as kind of local conditioner for learning another value. This sort of secondary parameter interaction was absent in the simpler 3-parameter space introduced in Section 2.2. This was because an equal number of grammars compatible with the input were V-final or T-final. Thus neither head-final value was more likely to be learned first, which is necessary for it to exert the kind of influence we see in secondary parameter interaction.

Although secondary parameter interaction favors a non-raising grammar in the 5-paramter space, it is still possible to learn a verb raising grammar, as attested in the results in Table 4.4. This can happen if V-final is not overly favored early on during the learning process. Given the randomness of sampling a compatible grammar, it is possible for, say, a *v*-final grammar to be most favored initially. There are 12 *v*-final grammars compatible with SOV input; given a *v*-final grammar a majority of the grammars (8 of 12) have raising to *v*.

The existence of secondary parameter interaction raises an important question for future research. Does enriching the model's parameter space or input result in sufficient parameter interaction such that a raising grammar cannot be learned? With respect to basic parameter interaction (i.e. not secondary parameter interaction), I have already discussed in Sections 3 and 4 how expanding the corpus and hypothesis space can still maintain the fundamental tension between raising and non-raising. But what about secondary parameter interaction? When we factor in the effect of secondary parameter interaction, the prospect of still being able to learn a raising grammar appears promising. First, there is no clear candidate for an additional parameter that would increase the effect of secondary parameter interaction that favors non-raising.[11] Examples of additional parameters are Object Raising (cf. Section 4.2), C-initial/final, and T-to-C movement. Including these parameters in the hypothesis space can increase the proportion of compatible grammars that are V-final (and sometimes slightly decreases this proportion). Second, with these additional parameters, when we consider compatible grammars that are V-final, the proportion of non-raising grammars does not increase from what we have seen above.[12] And as for additional types of input, I have already discussed how intransitive verbs with modifiers are a core example of how we might see secondary

[11] I set aside the case of parameters concerning the position of the subject. A more enriched model could contain various parameters of this sort, and these certainly have the potential to interact with the parameters that have been the focus of this study. Any possible interactions would of course depend on a precise formalization of these subject-related parameters, but at present such formalization appears to be the subject of much ongoing research in syntax. This is in part due to the number of positions subjects are claimed to occupy. Various analyses of the same language propose that sometimes the subject may remain in *v*P/VP, while at other times the subject moves to SpecTP or SpecCP. Examples of such analyses for a verb-final language (Japanese) may be found in Abe (1993) and Miyagawa (2001), and for German in Rosengren (2002) and Haider (2010). Accordingly, the step of including such parameters in the model should be taken tentatively, as it is a more speculative one. Nevertheless, a preliminary analysis of some possible candidates for these parameters (e.g. a parameter for raising the subject/highest nominal to SpecTP) indicates that the observations regarding parameter interaction and verb placement in this chapter extend more or less in a fairly general way to a model that includes these parameters. In this work, I have chosen instead to focus on what I consider to be clear-cut and fundamental parameters relating to verb placement, such as verb raising and head-complement directionality. I have shown on a small scale how these parameters can interact with each other and how such interaction plays an important role in the learning process. I leave further investigation of how subject placement fits into the larger picture of parameter interaction for future research.

[12] Interestingly, in the 8-parameter hypothesis space with V-to-C movement from note 2, there is a much stronger effect of secondary parameter interaction in favor of verb raising. An open question, then, is whether a non-raising grammar could be learned given such a hypothesis space.

parameter interaction that favors a non-raising grammar. In Section 3.2 we saw how this interaction was possible when the model received input with modifiers and had a parameter for leftward/rightward adjunction. However, as was discussed earlier, we expect the effect of this secondary parameter interaction to be relatively minimal given a more basic parameter interaction that favors a raising grammar. Thus the degree of secondary parameter interaction favoring non-raising does not significantly increase with these expansions to the model, and learning a raising grammar remains a viable option. It remains an open question, then, whether there are any clear ways of expanding the model that would contribute to a significant increase in the effect of secondary parameter interaction that favors a non-raising grammar.

Having discussed the model's results, we can now make a general comparison with the experimental findings in Han et al. (2007), henceforth HLM. Recall that HLM found that both adults and children split roughly into two groups concerning the scope of negation and a quantified object. Some participants consistently had high scope for negation, whereas other had low scope. HLM attributed this variability across speakers to speakers having different grammars concerning verb movement. Further, this variability was possible because the input to Korean learners does not provide sufficient unambiguous evidence for the structural position of the verb. Both verb raising and non-raising grammars are compatible analyses. These findings dovetail with the results from the model discussed above. I have shown how ambiguity in the input regarding the position of the verb persists in a multi-parameter model, and that there is insufficient parameter interaction to always push the model toward learning the same parameter setting for verb raising. This is borne out in the results where we see that the model variably learns a raising or non-raising grammar. In sum, the experimental findings in HLM's study can be taken as supporting evidence for the modeling results of variability.

At this point it is perhaps tempting to make a closer comparison of the proportions of the sub-populations in HLM's study and the proportions of raising/non-raising grammars in the models. Having discussed how secondary parameter interaction results in the model favoring a non-raising grammar, we might want to see whether such a preference can be found in the experimental results. Although this is a natural and potentially fruitful step for future research, as I discuss below, the limited nature of the results presently makes such a comparison rather premature.

By way of exposition, I first note HLM's brief comments about the distribution of sub-populations. Observing that the input is ambiguous for verb-raising parameters, HLM limit themselves to saying that when faced with such ambiguity, the learner randomly chooses a parameter setting. Such random selection is, of course, the basic learning mechanic of a probabilistic model, including the one proposed here. HLM do not elaborate further on how this random selection works precisely. They seem to suggest that selecting a parameter value is a 50-50 choice (perhaps like flipping a fair coin), because they refer to the split in the Korean population as being approximately 50-50 (i.e.

half have verb-raising grammars, and half have non-verb raising grammars). However, HLM do not present a formal learning model, and so it is not clear how such a split would be captured under their analysis. For example, one of the implications of this chapter is that an input-driven learner can have different results depending on the intricacies of parameter interaction. As HLM present no model of their own, it is not clear what effect parameter interaction would have in modeling their analysis.

When we look more closely, though, the results that HLM present do not clearly support such a 50-50 split in the population. Indeed, the picture that emerges does not particularly lend itself to clear extrapolation of the proportions of the sub-populations. The acceptance rate for wide scope of negation in Tables 4.2 and 4.3 is closer to one-third than one-half, and the adult and child populations in HLM's experiment split into groups that are closer to a 2:1 ratio than a 1:1 ratio. Further, the adults tested in Han et al. (to appear) have an average acceptance rate of 75% for wide scope of negation. I take these results to show that there is a real split in the population of Korean speakers, and I follow HLM in assuming that this is reflective of grammars that have different parameter values for verb raising. However, it seems premature to make any claims about the proportion of this split; a more precise characterization of the sub-populations awaits further experimentation with a larger sample size. At this early stage in experimental research, a goal for a learning model should be capturing some variability, though how much has yet to be determined.

5.2 Variability with a probabilistic learner: A broader perspective

In the literature, probabilistic learners have been used to model two other types of variability. To my knowledge, the learning scenario represented by the case at hand of verb raising has not been discussed before. In this section, I give an overview of the types of stable variability that have been covered, comparing them with the case of verb raising in Korean.

As was discussed in Chapter 2, Pearl (2007) considers the theoretical case of a single parameter being set solely on the basis of ambiguous input. When using a probabilistic learner such as Yang's (2002), Pearl shows that so long as the learner is not overly conservative, the learner can converge on either value of a binary parameter. In the context of this chapter, this type of stable variability can be characterized as variability driven by ambiguous evidence and lack of parameter interaction. However, the lack of parameter interaction is simply an artifact of considering a single parameter in isolation. What has been demonstrated in this chapter is that stable variability is also possible in a multi-parameter system given ambiguous input and insufficient parameter interaction. This chapter has also looked at some of the necessary conditions for symmetry between a parameter's values, so as to arrive at stable variability in such a multi-parameter system

Pearl (2007) also models a different type of variability, which is found within speakers of a language undergoing diachronic change. Pearl looks at the case of word

order in Old English. For example, in the transition from Old English to Middle English, there was a stage when OV and VO sequences were possible. This variability is not stable in that the same author of a text makes use of competing grammars (cf. Pintzuk 2002). In terms of a probabilistic model, this can be captured by assuming that the adult learner has converged on a state where a parameter's probability mass is not concentrated primarily on one value, but rather is distributed more equally across all the relevant parameter's values. For Old English learners, some non-trivial proportion of the probability mass would be placed on, say, V-initial, while the remainder would be placed on V-final. Although it is often proposed that this kind of diachronic change is initiated by ambiguous input (e.g. Roberts 1997), in subsequent stages of the diachronic change, this diachronic grammatical variability is claimed (in part) to result from a learner receiving multiple kinds of unambiguous evidence that are at odds with each other. The Old English learner, then, learns from input that contains unambiguously OV utterances as well as unambiguously VO utterances. This kind of variability, which is found in these subsequent diachronic stages, is what Pearl has modeled. However, this kind of variability contrasts with the variability exemplified by verb raising in this chapter. Unlike this diachronic example, with verb raising in Korean there was no input that forced the learner to adopt a raising grammar or a non-raising grammar. Rather it was the interplay of parameters in the face of ambiguous input that resulted in conditions where variability could take place.

The case of stable variability in this chapter thus adds a new type of syntactic variability to the emerging modeling literature on this topic. This chapter has highlighted the existence of this type of variability in a multi-parameter system and has illustrated some of the conditions necessary for it to be possible.

6. Comparison with other models

I now briefly compare the model proposed here with the alternative learning models introduced in Chapter 2. This section reviews some of the main points made in Section 3 of Chapter 2, recasting them in terms of the verb-final/Korean discussion of this chapter. Unlike the probabilistic model proposed here, some of the models struggle to systematically learn head-final values for multiple projections given ambiguous input. This can be remedied by appealing to default values, but the cost of this is not being able to model variability across speakers with respect to verb raising. In contrast, the probabilistic model of Yang (2002) is well suited to address the learning puzzles we have seen in this chapter. Nevertheless, the role of parameter interaction in learning from ambiguous input is a way of learning that is not explored in any detail in Yang's work. A contribution of this chapter, then, is to shine a light on a previously untapped possibility for learning syntactic parameters.

An input corpus that is fully ambiguous for headedness and verb raising, as I have assumed Korean is, presents a learning challenge for a deterministic learning model such

as Sakas and Fodor (2001), which explicitly learns only from unambiguous input. Such a model seemingly has no way of learning basic properties of phrase structure from the evidence available. One possibility available to this kind of model is to adopt default values for parameters. Moreover these defaults would constitute the end-state of learning. That is, the learner would not be able to learn any parameter settings on the basis of these defaults. To see why, recall that all the input is ambiguous for any given parameter value. If the initial state involved some set of parameters that was not compatible with the input, the learner would not be able to leave that state because of the poverty of unambiguous evidence for any particular parameter value. Crucially, the interaction of the parameters here is such that setting one always depends on the value of another. However, as a heuristic approach of deterministic learning, the learner cannot rely on a default value to set another parameter without first having seen unambiguous evidence that the default value is correct (and such evidence, by hypothesis, does not exist in Korean, or is vanishingly rare).[13] Were the learner to rely on an unverified default value, they might make a mistake from which it would be impossible to recover (cf. Fodor 1998; Sakas and Fodor 2012). Thus even when equipped with a set of defaults, a deterministic learner could not learn to set any parameter values that differed from the initial state. Accordingly, to ensure that all learners had an initial state that was compatible with Korean, there would need to be a consistent set of default values for all learners. Regardless of what this Korean-compatible initial state is, though, it would not lead to variability across learners. For example, to achieve the same results as the probabilistic model, these defaults would need to be head-final for all the head-directionality parameters. As the learner never receives evidence that is incompatible with these head-final defaults, converging on a consistently head-final grammar is guaranteed. However, this approach would also require some default setting for verb raising. The problem with such a default, though, is that it predicts that all speakers will have either a raising or a non-raising grammar: with no unambiguous evidence in the corpus in (28) to push the learner from the default setting we no longer predict variability across speakers. We thus have no way of modeling the variability reported in Han et al. (2007) for Korean.

In contrast, the Triggering Learning Algorithm of Gibson and Wexler (1994) is able to capture this variability. There are multiple target grammars the learner could converge on, some of which are raising grammars and some of which are not. Yet it this very same flexibility that allows for more learning outcomes than predicted by the probabilistic model in this chapter. Consider the range of grammars compatible with SOV input from the original example in (10), repeated below:

[13] Recall from Section 4.1 that I follow Han et al. (2007) in assuming that the relevant evidence that unambiguously distinguishes the scope of negation is sufficiently rare in the input to the learner that it has little to no role in the average learner's development.

(10) *Grammars compatible with SOV*
a. [+VT, *T-fin*, V-in] c. [–VT, T-in, V-fin]
b. [+VT, *T-fin*, V-fin] d. [–VT, *T-fin*, V-fin]

Any of these grammars could be the initial state for the learner; if the learner starts with one of these grammars, that will also be the end-state for learner because there is no input that is incompatible with one of these grammars. Recall that the learner will only adopt another grammar when forced to by input incompatible with the current grammar. (10) indeed includes both raising and non-raising grammars, but this set of grammars also includes both head-initial and head-final grammars. The TLA learning model thus has no principled way of learning head-final values for multiple projections by taking advantage of the probabilistic tendencies of parameter interaction. This is a greater range of variability than we saw in either the modeling results in this chapter or in the experimental results of Han et al. (2007). As discussed in Chapter 2, this is not a problem *per se* for the TLA. Both my model and the TLA provide evidence against the null hypothesis, according to which there is no variability across speakers. The difference is that currently only variability for verb movement finds additional empirical support (cf. Han et al. 2007). The balance of evidence available is thus currently in favor of the approach in this chapter, although the TLA's greater range of variability could find further support in future experimental work. Again, the TLA could make use of default parameter values in an attempt to constrain the range of variability, a possibility that Gibson and Wexler discuss. The model could then have head-final default values for the head-directionality parameters. With head-final defaults, regardless of what the default value for verb raising is, the learner would never adopt a grammar different from the that of the initial state: all head-final grammars are compatible with the input under either a raising or non-raising parameter setting. We are now in the same problematic situation as above when amending the model of Sakas and Fodor: in an attempt to learn head-finality for multiple projections in a principled way, the model is no longer able to reflect grammatical variability across speakers.

Finally, it is important to stress that a probabilistic learning model such as Yang's (2002) is fully compatible with the results of the model reported here. What I would like to highlight here is an extension of the point made in Chapter 3 with respect to Yang's model: the work here serves to shine a light on the role of parameter interaction in the learning process. Indeed, the model in this chapter is highly comparable to Yang's in its basic approach to learning: by abstracting away from a fully generative model that outputs sentences, the model here simply looked at the compatibility of the input, which is the essence of Yang's model. Thus Yang's model is also predicted to consistently learn a grammar that has multiple head-final projections with variability for verb raising via the same basic mechanism of parameter interaction discussed in this chapter. However, this is a point that is largely overlooked in Yang's discussion of learning from ambiguous input. Recall from Chapter 2 that one of Yang's core examples of learning from

ambiguous input involved learning a V2 grammar. Yang observed that all the input for a V2 grammar was compatible with some other grammar. Thus all the input for a learner of a V2 grammars is ambiguous. Crucially, though, none of the grammars competing with the V2 grammar were compatible with all of the input for the V2 grammar. It was simply not necessary to investigate parameter interaction in such a scenario because all the competing grammars were always punished by a certain kind of input. In the example of a verb-final language presented in this chapter, there are multiple competing grammars that are all compatible with all the input. Parameter interaction now plays a vital role in distinguishing which grammar is more likely to be learned. This, then, is the sort of example that brings the contribution of parameter interaction to the forefront in a discussion of learning parameter settings.

7. Constraining the model: A first attempt

In Chapters 3 and 4, we have seen some of the successes of the model in accounting for learner errors and variability in Swiss German and Korean. A question now facing the model is whether it needs to be constrained in some way. That is, might the model overgeneralize by learning a grammar that is not attested cross-linguistically, and if so, should the model be amended so as to preclude such overgeneralization? This relates to the question of language universals. Suppose there is some gap in the inventory of the grammars of the world's languages. One possibility is that this gap is not accidental and that it exists because the unattested grammar very hard to learn (possibly unlearnable).[14] From this perspective, a possible criterion for the overall success of the model is how successful it is in not being able to learn the unattested grammar.

The purpose of this section is to investigate some of the difficulty involved in constraining the model. The challenge will be to constrain the model so that we obtain the kind of modeling results we have already seen while still making an unattested end-state an unlikely outcome more generally. This discussion will hopefully set the stage for future research that can more fully address this challenge. A preliminary investigation in begun in Chapter 6, where I explore the possibility that learning biases can help to effectively constrain the model.

In this section, I illustrate this challenge by focusing on an example of a putative language universal that is particularly relevant to the kinds of parameters I have been investigating, namely the Final-over-Final Constraint (FOFC) of Biberauer et al. (2014), which I have alluded to at various points in the discussion. The descriptive generalization of FOFC is that certain combinations of values for parameters of head-complement order

[14] Whether or not such an unattested grammar is unlearnable is ultimately an empirical question, one that can perhaps be addressed with artificial language learning experiments. To my knowledge, this question has not been investigated. Consequently, it seems too strong to me to claim that such a grammar is unattested because it must be unlearnable as opposed to claiming that the grammar must be very hard to learn (thereby being very unlikely to occur).

are unattested cross-linguistically. As we will see, nothing in the current implementation of the model constrains it from reaching an end-state that runs counter to this universal. The model can learn a FOFC-violating grammar as an end-state because learning one parameter value does not intrinsically bias the learner toward learning some other parameter value. The challenge will be to still capture the results concerning verb movement in Korean while making unlikely end-states that are FOFC-violating when learning other languages. I illustrate the challenge here with a first attempt at so constraining the model. Although this attempt is successful in ruling out an end-state that violates the universal, it does not predict the kind of variability we have seen in Korean. At this point I must leave the proper treatment of how to constrain the model with respect to FOFC as a topic for future research, but I present a more speculative proposal of how to do so in the broader discussion on learning biases in Chapter 6.

The issue of language universals also relates to the notion of principles in a Principles and Parameters framework. In this thesis I have focused on the role of parameters in the learning process, and we have seen how this approach has had its successes. This is not to say that principles should be overlooked. As principles represent some invariant property of grammars across languages, they naturally relate to the issue of systematic gaps in grammars cross-linguistically. Again, a plausible approach to these gaps is that they are due to some principle of grammar, and again the question becomes how to implement such a principle in the learning model without losing the results we have already obtained. If some principle of grammar is ultimately behind the generalization of FOFC, we can thus use the discussion on FOFC as an illustrative example of the care that should be taken in considering how to implement a principle in the learning model. In Chapter 6 I speculate that this kind of principle can be incorporated into the model in the form of a learning bias.

In the remainder of this section I first present a general introduction of an example of a putative language universal, namely the Final-over-Final constraint (FOFC) from Biberauer et al. (2014). I then illustrate the difficulty in capturing the variability we have seen in Korean while constraining the model from reaching an end-state that is inconsistent with this universal.

Recent work such as Biberauer et al. (2014) claims that the following structural configuration is unattested in the languages of the world: within certain domains, such as the extended projection of the verb, a head-final phrase does not take a head-initial phrase as its complement. This generalization is given the name the Final-over-Final Constraint, or FOFC. The claim then is that the structures in (43) are unattested; we can call such structures FOFC-violating. (43a) gives the constraint schematically, and (43b) provides one example of a FOFC-violating structure, which involves the phrases VP, TP, and CP. The part of the structure of (43b) that is FOFC-violating (in boldface) is the relationship between VP and TP: a head-final TP takes a head-initial VP as its complement.

(43) *Unattested grammars cross-linguistically as per FOFC*
a. [$_{XP}$ … [$_{YP}$ … Y ZP] X …]
b. [$_{CP}$ C [$_{TP}$ [$_{VP}$ **V O**] **T**]]

Let us suppose that, indeed, these word orders are unattested (though see Biberauer et al. for discussion of numerous potential counterexamples). Further, along the lines of the discussion above, let us assume that the orders are unattested because they are very hard to learn, if not impossible. If these word orders are hard to learn, then we can consider the following desideratum for the learning model. Even when a hypothetical input corpus contains input that is only compatible with a FOFC-violating grammar (e.g. the string […Comp Verb Object Tense]; cf. (43b)), the model would still have difficulty in reaching an end-state with a grammar that is FOFC-violating. In the simulations of Swiss German and Korean, we have seen how the model can always learn a grammar that is FOFC-compatible. However, in its current implementation there is nothing about the model that impedes it from reaching a FOFC-violating end-state. If a token of input is only compatible with a FOFC-violating grammar, the model must currently sample and reinforce parameter values that are good fit to this input. In no sense does the model constrain, say, sampling a grammar with a head-final TP and a head-initial VP: any parameter value can be sampled and reinforced so long as it is consistent with the input. Given sufficient input of this sort, the model will be pushed toward a FOFC-violating end-state.

How, then, can we constrain the model from ending up with a FOFC-violating grammar? As a first attempt at addressing this question, I make the following proposal in this section: FOFC-violating grammars are not considered by the learner during the learning process. If the learner never considers a FOFC-violating grammar, then it follows that the end-state will not be a FOFC-violating grammar. Such a proposal would seem to satisfy the desideratum above by avoiding a FOFC-violating end-state, but in what follows I point out several difficulties for this proposal.

First, it is actually not entirely clear how such a proposal would be implemented in the learning model I have proposed. As noted in Chapter 1, the learner's hypothesis space is not populated with an enumeration of grammars and their respective weights, but rather with a set a parameter values and respective weights. Thus enforcing FOFC is not simply a matter of eliminating certain grammars from the hypothesis space. There are no grammars in the hypothesis space.

Moreover, even if we do try to implement this proposal, I want to show that it can have an unwelcome consequence in modeling variability across learners. To that end, let us simply adopt a brute-force filter that can enforce FOFC. Assume that the learner has a list that enumerates all FOFC-violating grammars, and that the learner can use this list to enforce FOFC. Let us now walk through the learning procedure. (As an example, I will use the simplified version of the model as I did earlier in this chapter; nothing crucially hinges on this, and the discussion would be similar under the fully generative version of

the model from Chapter 2.) The learner receives a token of input and constructs a grammar by sampling a value from all the parameters. If this sampled grammar is incompatible with the input, then learner samples again until a grammar that is compatible with the input has been constructed. The values of this input-compatible grammar are then reinforced. Consider now what happens if this input-compatible grammar is on the list of FOFC violators. The learner can enforce FOFC by discarding this grammar. In other words, the learner will continue to construct a grammar until the grammar sampled is both (a) compatible with the input; and (b) does not violate FOFC. This has the effect of essentially removing certain grammars from being considered as viable targets. This is an admittedly *ad hoc* filter, but it will do the trick to illustrate the point at hand.[15, 16]

Let us now revisit the case of Korean from this chapter to consider which grammars are FOFC-compatible. I will illustrate this for the 5-parameter implementation of the model, but a similar point can be made for the simpler 3-parameter version. Again, we are interested in which grammars are compatible with SOV input. Recall that it was only with SOV input that some grammars were incompatible; all grammars were taken to be compatible with SV input. Now we are not only interested in which grammars are compatible with the input, but also what subset of these grammars do not violate FOFC. This is illustrated in (44) below. All the grammars in (44) are compatible with SOV input. On the left in (44a) are grammars that are also FOFC-compatible. The FOFC violating (but input-compatible) grammars are on the right in (44b).

[15] In light of the discussion from earlier in this section, this filter raises the question of what the learner does when the input is only compatible with a FOFC-violating grammar. Although this is an important question for a filtering approach, it is orthogonal to the point I am making in the main text: even when given an input corpus such as Korean, which does not contain such input (i.e. input that is only compatible with a FOFC-violating grammar), this filtering approach does not allow us to capture the variability that was modeled earlier in this chapter. Consequently, I will not attempt to propose an answer to this question for the filtering approach.

[16] As far as I can tell, this filter makes the same predictions as Biberauer et al.'s (2014) proposal to account for FOFC. Their proposal appears to use syntactic features to essentially stipulate that certain structures are ruled out by UG.

(44) *Grammars compatible with SOV input from the 5-parameter model*

a. FOFC-compatible	b. FOFC-violating
1. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]	1. [+V-*v*, +*v*-T, T-fin, *v*-in, V-in]
	2. [+V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
	3. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-in]
2. [+V-*v*, –*v*-T, T-in, *v*-fin, V-fin]	4. [+V-*v*, –*v*-T, T-in, *v*-fin, V-in]
3. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]	5. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-in]
4. [–V-*v*, +*v*-T, T-in, *v*-in, V-fin]	6. [–V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
5. [–V-*v*, +*v*-T, T-in, *v*-fin, V-fin]	
6. [–V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]	
7. [–V-*v*, –*v*-T, T-in, *v*-in, V-fin]	7. [–V-*v*, –*v*-T, T-fin, *v*-in, V-fin]
8. [–V-*v*, –*v*-T, T-in, *v*-fin, V-fin]	
9. [–V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]	

We see that our *ad hoc* FOFC-filter has knocked out almost half of the grammars. This means that the learner will only sample and reinforce parameter values from the grammars in (44a). Which values do we expect to the model to learn? Crucially, we are interested in raising the verb to *v*P. We see in the partial summary in (45) that a healthy two-thirds majority of the grammars do not have V-to-*v* movement, i.e. they are [–V*v*]. In other words, we have lost the symmetry between verb raising and non-verb raising grammars that we saw earlier, in which half of the grammars the model could sample had verb movement and half did not.

(45) *Partial summary of grammars in (44a)*
a. +V-*v* grammars: 33.33%
b. –V-*v* grammars: 66.66%

We also saw earlier in this chapter that with the same kind of input corpus, a two-thirds majority of grammars that favors a particular parameter value is sufficient to push the learner toward a parameter setting of that value. The summary in (45) thus leads to the following conclusion. If we were to run a simulation of the model that included the FOFC-filter, we would not expect the model to ever learn a verb raising grammar. Enforcing FOFC, then, is expected to have the unwelcome effect of losing our account of modeling grammatical variability across learners. In this chapter we have seen that such variability was possible precisely because the model could reinforce parameter values when sampling the FOFC-violating grammars in (44b).

I conclude that the case of variability in Korean can be taken as evidence against the proposal to eliminate FOFC-violating grammars from the learner's consideration. But if sampling a FOFC-violating grammar is possible during the learning process, we are left with the question from above of how to constrain the model from reaching an end-state

that is FOFC-violating.[17] In Chapter 6 I make a different sort of proposal, albeit a more speculative one, in an attempt to address this question. There I propose that learning biases concerning how parameter values are reinforced might prevent the model from reaching a FOFC-violating end-state. What I hope to have illustrated in this section is that there is clearly more work to be done with respect to constraining the learning model and that part of the challenge is doing so in a way that still captures the modeling results I have presented. Nevertheless, the current absence of a clear-cut way of successfully constraining the model should not be taken as a reason to peremptorily discard the model. The model has given us a principled way of accounting for various empirical phenomena such that they fall out from the learning process itself. The hope is that questions such as how to constrain the model stimulate future research that builds on these results in an attempt to also account for language universals.

8. Summary

In this chapter I have shone a light on the possibility of learning systematically from ambiguous input in a multi-parameter space by closely looking at the effects of parameter interaction. We have seen that even when unambiguous evidence is entirely lacking, parameter interactions can lead the learner to systematically learn head-finality for a particular set of syntactic projections. I presented a simple model of this for the verb-final language Korean. Moreover, this model is able to capture the effect of variability in a population when there is a high degree of ambiguous input and a lack of sufficient parameter interaction. I modeled this variability with verb raising and attempted to draw a close parallel between the results of the model and experimental results in the literature that are also claimed as evidence for variability in verb raising in the Korean population at large.

[17] Another question relates to child productions during the course of acquisition that reflect a FOFC-violating grammar. In the model the learner can sample a FOFC-violating grammar, and so we might expect the learner to sometimes adopt a FOFC-violating grammar during acquisition. The expectation for acquisition is that children might sometimes make errors that contain FOFC-violating structures. That children might produce utterances that reflect grammars that are unattested among adult grammars is certainly plausible if we follow, for example, Wexler's (1998) account of the Optional Infinitive stage. Do children ever produce utterances that contain FOFC-violating structures? I am not aware of any attested examples, but I think this should be left as a question to stimulate future research. I believe the typology of child errors is simply not rich enough to deny the existence of a phenomenon that might be relatively rare. Indeed, early parameter setting for some parameter values can eliminate FOFC-violating structures from appearing. For example, if there is robust evidence that V is head-final, the FOFC-violating structure in (43b) becomes the following FOFC-compatible structure: [$_{CP}$ C [$_{TP}$ [$_{VP}$ **O V**] **T**]].

Chapter 5

The Case of Zero-Derived Causatives in English: Learning from implicit negative evidence

1. Introduction

The previous two chapters have looked at the interplay between ambiguous evidence and parameter interaction. In this chapter, I look at a special case of learning from ambiguous evidence, namely the subset scenario. The learning challenge of the subset scenario is as follows. If the target grammar is the grammar of a subset language, how does the learner learn this more restrictive grammar when the evidence available to the learner is consistent with both that grammar and the grammar of a superset language? That is, the learner's input underdetermines the correct structural analysis. Further, the difference between these two competing grammars can be reduced to a single parameter setting that does not obviously interact with any other parameters. Thus parameter interaction will not be of any help in learning the grammar of the subset language, as only one relevant parameter is at play. The subset case, then, presents another ideal testing ground for the probabilistic learner I have proposed. All the relevant input is ambiguous, and parameter interaction does not play a role. I propose that the fully generative model introduced in this thesis can address the learning challenge by learning from implicit negative evidence. In the model, the learner has different expectations concerning the shape of the input under the grammars of the subset and superset languages. The learner is thus sensitive to not hearing evidence that would more strongly support the grammar of the superset language. As all the input is from the grammar of the subset language, it is this grammar that is the grammar of best fit, and that is what the model will learn.

The case study under consideration involves learning the syntactic structure of zero-derived causatives (ZDCs) in English. Following Pylkkänen (2008), the structure of causatives hinges on setting the Cause-selection parameter. Different values result in syntactic structures of different sizes, or complexity. The simpler structure represents the grammar of a subset language, whereas the more complex structure represents the grammar of a superset language. As is discussed more fully in the following section, Pylkkänen's claim that the ZDC in (1) instantiates the more restrictive grammar, with the simpler syntactic structure of the subset language, is based on negative evidence involving modification with the manner interpretation of adverbs. Thus the interpretation of (1) is that John's action can be characterized by grumpiness, and not Bill's awakening.

(1) John awoke Bill grumpily/happily.

According to Pylkkänen, the fact that these examples mean one thing and not the other is evidence for a particular parameter setting. Languages with a different parameter setting would allow different interpretive possibilities for an example like (1). The claim based on (1) is thus based on negative evidence, the absence of a particular interpretation. Additionally, there is no positive evidence from English that uniquely identifies any other value for the Cause-selection parameter as being correct/incorrect. In fact, as we will see, the positive input data relevant to setting the parameter actually underdetermine the correct structural analysis. That is, all the input data is ambiguous. Given this, the challenge is how the correct parameter setting can be learned.

The learning challenge can also be framed in terms of the Poverty of the Stimulus (cf. Chomsky 1980). Given that the learner receives no overt evidence pertaining to the correct parameter setting (i.e. the grammar of the subset language), the question is then whether this parameter setting is unlearnable. The approach taken here is that in the type of learning situation exemplified by (1) the target grammar is indeed learnable. And the claim is that it is learnable by means of implicit negative evidence.

The primary goal of this chapter, then, is to present of proof-of-concept illustration of how the learning model can address the learning challenge by consistently learning the grammar of the subset language. The model does this by learning from implicit negative evidence, which in the case of the causative in (1) involves learning from the absence of the unattested interpretation. We will see that the model is highly successful when learning from a simple corpus of the sort in (1). Given that we see success at this basic level, the hope is that scaling up the model in the future will also be a viable approach for subsequent research.

This chapter contributes another example of the role that ambiguous evidence plays in the learning process. In particular, this chapter provides a clear empirical example of the subset scenario. Although the challenge of learning the grammar of a subset language has figured prominently in the learnability and acquisition literature (e.g. Gold 1967; Wexler and Manzini 1987), more recent work has questioned whether examples of the subset scenario are attested in natural language within the domain of syntactic parameters (cf. Atkinson 2001; Frank and Kapur 1996; Hyams 1986). The current chapter thus serves to renew this line of inquiry from a new perspective, as well as to bring Pylkkänen's theory of causatives more generally into the domain of acquisition.

I note that although I will be following the theoretical framework of Pylkkänen in many respects, the challenge of learning the parameter setting of the grammar of the subset language is one that could emerge in various other theories that assume that ZDCs are structurally complex (cf. Hale and Keyser 2002; Ramchand 2008; and Folli and Harley 2005). In Section 2, I consider a variant of Pylkkänen's theory and show how the learning challenge remains essentially the same. Thus the learning challenge is not intrinsically tied to Pylkkänen's theory, but I will use Pylkkänen's framework as a clear way of illustrating what the challenge is.

As will be discussed in more detail, the correct parameter setting for ZDCs in English is the simpler structure, and a second contribution of this chapter is to show that this can in fact be learned. In other words, the choice of the simpler hypothesis over a more complex one is something that is derived by the actual learning procedure, and there is no need to invoke some principle such as the Subset Principle (Berwick 1986), or to resort to default values for parameter setting.

Another advantage of the model proposed here is that it has success with the learning challenge, whereas other models do not. I review several other models, both probabilistic and non-probabilistic, and show how they are not up to the learning challenge as they are currently formulated because they do not learn from implicit negative evidence.

Finally, this chapter builds on earlier work involving learning from implicit negative evidence. I do this by pulling together several different strands of research. First, as a general approach, the work here pursues a direction of research from at least Braine (1971) that focuses on how implicit negative evidence can assist the learner (see also Bowerman 1988). In more recent work, the role of learning from implicit negative evidence has been developed in hierarchical generative Bayesian learning models, such as Hsu and Griffiths (2009) and Perfors et al. (2010). The full version of the model I have proposed is a generative model in a similar vein.

What these earlier models, such as Perfors et al. (2010), do not do is take a detailed look at learning syntactic structure. A second line of research pursued here brings to the discussion the kind of enriched phrase structure that is learned in the probabilistic context-free grammar of Perfors et al. (2006). Indeed, as surveyed in Pearl and Goldwater (in press), much of the emerging body of literature using Bayesian modeling methods in generative linguistics has not looked at syntactic parameter setting. The work in this thesis builds on this earlier work by incorporating both learning from implicit negative evidence and the learning of syntactic structure.

As was mentioned in Chapter 1, this work is in some respects most similar to the line of research concerning anaphoric *one* in English, initiated in Regier and Gahl (2004) and developed in a series of papers by Lisa Pearl (Pearl and Lidz 2009; Pearl and Mis 2011; Pearl and Mis in press). These works look at the intersection of implicit negative evidence and learning from ambiguous input. Importantly, these works also look at learning the syntactic structure of *one*. More recent work, such as Pearl and Mis (2011) and Pearl and Mis (in press), has also developed the proposal using a hierarchical model. However, there are a number of differences between them and the current study. Perhaps most significantly for the discussion here is the recognition that indirect positive evidence plays a crucial role in learning anaphoric *one* (e.g. Pearl and Mis 2011). Further, as Payne et al. (2013) note, (a) not all input the learner receives concerning anaphoric *one* is ambiguous (though the unambiguous evidence may be vanishingly rare; cf. Pearl and Mis

in press), and (b) the properties that the model attempts to learn reflect only preferences in the adult grammar.

In contrast, we will see that implicit negative evidence is sufficient, within the scope of the model here, for addressing the learning challenge with zero-derived causatives. Thus with the simple corpus of input that the model learns from in Section 3, the model can capitalize on implicit negative evidence to arrive at the grammar of the subset language. Future research can look at what the limits of this success are when the corpus is expanded (see the discussion in Section 3.4; cf. similar discussion in Lidz and Pearl 2009). Moreover, in the case of ZDCs, there is no unambiguous evidence, and Pylkkänen's claim is that the parameter setting is categorical (i.e. not a preference). Given these differences, the present study is an ideal case that serves to shine a light on the crucial role of implicit negative evidence in parameter setting for the learner.

In the discussion we will encounter data that appear to be somewhat rare in the primary linguistic data. One might wonder whether this data is too infrequent for children to learn from. To address this, I try to find alternative routes for learning that do not rely on the possibly rare data. As a general heuristic, then, I will pursue an approach to language acquisition that learns as much as possible from common, relatively frequent input. Indeed, this was part of the approach in Chapter 4, where I assumed that potentially unambiguous data involving negation was simply too rare in the learner's input to play much of a role in parameter setting. To this end, I propose two ways in which the learning model might be implemented. The first focuses on learning from implicit negative evidence when the input contains modifiers as in (1). There is reason to think that such input is relatively rare. However, the learning mechanism is more general and can be extended to learn in the same way from non-modified input, as in (2), which is much more commonly attested. Non-modified input such as (2) is also ambiguous between the grammars of the subset and superset languages. Nevertheless, the generative model can learn the simpler syntactic structure of the subset language.

(2) John awoke Bill.

To the extent that frequent input is sufficient for language learning, it takes the question of learning from rare data out of the spotlight and puts it to the side. Further, it leads to interesting considerations, such as what kind of languages the learning model predicts to be or not be possible (Section 3.4).

The structure of the chapter is as follows. In Section 2, I review the core aspects of Pylkkänen's theory of causatives and introduce the learning challenge it presents for zero-derived causatives in English. I then discuss in Section 3 how the model can learn the simpler structure via implicit negative evidence. This proposal is given in the form of a probabilistic generative model, which was introduced in Chapter 2. I give two implementations of the model I propose for learning the correct parameter setting for zero-derived causatives in English. The first in Section 3.1 is closely tailored to the actual

data Pylkkänen discusses. The second, revised implementation in Section 3.2, though similar to the first, is a more general extension, such that the first version of the model can really be thought of as a specific sub-case of the more general second version. Results of running this revised implementation are given in Section 3.3. Finally in Section 4, I discuss several other kinds of learning models that are not able to learn the correct parameter setting because they have no way of learning from implicit negative evidence.

2. Pylkkänen (2008) and the learning challenge

In this section I introduce the learning challenge presented by Pylkkänen's theory of causatives. In Section 2.1 I present an overview of Pylkkänen's theory and focus on the parameter that determines the complement of the Cause-head in zero-derived causatives in English, such as *break* or *melt*. There are different parametric choices that vary in structural complexity, and holding other parameter values constant, the languages that result from these choices are in a subset/superset relationship with each other. An analysis shows that in the target grammar, adults have a parameter setting for the simplest syntactic structure, namely a Root-selecting parameter setting. Thus the target grammar is the grammar of the subset language. The evidence comes from the unambiguous interpretation of adverbial modification of the causatives; Pylkkänen argues that the unambiguous interpretation is derived by the simpler structure but is left unexplained with a more complex structure. This argument relies on negative evidence, but leaves unanswered the question of how children could *learn* which parameter is correct. How do children learn that the simpler syntactic structure (i.e. the grammar of the subset language) is in fact the correct structure? As will be shown in Section 2.2, there is a learning challenge because all the evidence children hear underdetermines which parameter setting is correct. In other words, there is no positive evidence that is clearly in favor of one parameter setting over another. This is the challenge of the subset learning scenario. In Section 3 I show how the learning model addresses this learning challenge.

2.1 Review of Pylkkänen (2008)

In this section I present an overview of Pylkkänen (2008), focusing on the parameters that will be relevant for subsequent sections. I begin by reviewing Pylkkänen's evidence for zero-derived causatives in English being Root-selecting causatives.

The key evidence that Pylkkänen gives for causatives such as *break* or *awake* being Root-selecting involves adverbial modification.[1] Recall the example in (1) with a manner interpretation of the adverb, which is repeated below.

[1] Pylkkänen claims that additional evidence that zero-derived causatives in English are Root-selecting comes from the relation between unergative verbs and these causatives. I do not accept this claim and

(1) John awoke Bill grumpily/happily.

Pylkkänen's observation, (based on Fodor 1970), is that in examples such as (1), the adverb unambiguously modifies an event description involving the *causer* and not the *causee*. Thus in (1), John's action can be characterized by grumpiness, and not Bill's awakening. Pylkkänen notes, though, that the modifier in (1) is perfectly able to modify the event of Bill's awakening in the inchoative example in (3). See Vecchiato (2011) for various other examples that pattern like (1).

(3) Bill awoke grumpily/happily.

The question Pylkkänen asks is: if we follow Parsons (1990) in assuming that causatives involve two eventualities – a causing eventuality and a caused eventuality (cf. Dowty 1979) – why is it that the adverb in (1) unambiguously modifies the causing eventuality and not the caused eventuality?[2] We can say that modification of the causing eventuality gives us a high reading (e.g. the possible adverbial interpretation in (1)), whereas modification of the caused eventuality gives us a low reading (e.g. the impossible interpretation in (1)). Pylkkänen concludes that the lack of a low reading in (1) is due to a structural property of the causatives (their being Root-selecting; see below). If we follow Pylkkänen, then with respect to learning we can ask how the learner learns this structural property, such that only the high reading is possible in (1), a question I repeat in Section 2.2. In presenting Pylkkänen's argument as to why the low reading is not available in (1), I first introduce some of Pylkkänen's theoretical assumptions about the syntax and semantics of causatives.

Regarding the syntax of causatives, Pylkkänen assumes that there is a Cause-head in UG (which is phonologically null in zero-derived causatives) and claims that there is parametric variation as to what the complement of the Cause-head is. This is the Cause-selection Parameter, and Pylkkänen presents a three-way parameter for what the complement of the Cause-head can be. The different choices involve progressively more complex structures for the complement, and each more complex structure properly contains the structure(s) of the simpler structure(s). By hypothesis, this parameter is set independently of any other parameter setting. That is, the Cause-selection parameter does not interact with any known syntactic parameter. Further, this parameter is set relative to

---

discuss this in more detail in Appendix 2. Thus, I assume that the only evidence relevant for determining the correct parameter setting in English involves modification.

[2] Thus I follow Pylkkänen and much recent literature (e.g Ernst 2002) in assuming that such adverbs can modify eventualities. That is, these adverbs modify phrase markers along the clausal spine. Under this assumption, any analysis of (1) and (3) will not have recourse to some account that links the interpretations of these examples to some constraint concerning the relation between adverbs and grammatical roles. An example of such an account might be that manner adverbs can modify only subjects, and that the low reading in (1) could only arise via modification of the object, which would violate this constraint.

a particular causative morpheme. Thus a language with multiple causative morphemes could have the parameter set differently depending on the morpheme.

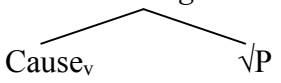
I will give two, similar versions of how this parameter can be formulated. The first is Pylkkänen's own, in which the structurally simplest complement is the projection of a category-neutral root. That these roots are in fact category-neutral is not clear, and the second variant is minimally different in assuming that the lexical root carries some category feature such as V. The analysis below of (1) is sufficiently general to be applied similarly under both variants. Thus both variants of the parameter will offer a comparable learning challenge. Considering such a variant helps to illustrate that the model's learning from implicit negative evidence is not tailored to fit Pylkkänen's theory, but is amenable to any theory of causatives that has competing parameter values that can generate languages that are in a subset/superset relationship.

In Pylkkänen's theory, the complement of the Cause-head could be a RootP (Root-selecting causatives); a *v*P (Verb-selecting causatives); or a θ_{Ext}P (Phase-selecting causatives). A Phase-selecting causative would be a causative morpheme that takes as its complement any projection that introduces an external argument. For the purposes of chapter, I will only consider Root-selecting and Verb-selecting causatives. This will allow for a more focused discussion on zero-derived causatives in English, as well as a more concise discussion of the learning challenge for parameter setting. The learning proposal I introduce in Section 3 could be scaled-up to accommodate all three parameter choices, but I leave this for future research.

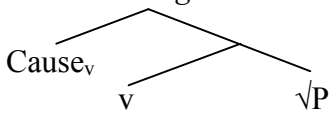
Schematic structures of Root-selecting and Verb-selecting causatives are given in (4).

(4) *Cause-selection (Variant I)*

a. *Root-selecting Cause*



b. *Verb-selecting Cause*



In this variant of the parameter, both structures have a category neutral lexical root that is embedded by the Cause-head. (See Marantz 1997 and Borer 2005 for discussion of category-neutral roots and category-defining morphology.) For zero-derived lexical causatives in English, this root could be √BREAK or √MELT and will be verbalized by a category-defining head. Before verbalization, though, the root combines with the internal argument and projects a √P.

The difference in the Cause-selection Parameter in (4) can be thought of as a difference in what functional head verbalizes the √P. Is it simply a category-defining little v^0 with no apparent (or necessary) semantic contribution, or is it the Cause-head, which is a flavor of little v^0 itself? The difference might appear to be slight, but a Verb-selecting parameter setting crucially results in a more permissive grammar, allowing for more

modification possibilities (as well as verbal morphology between the Cause-head and the root; cf. Section 3.4). Pylkkänen uses this difference in complement size for the Cause-head to account for cross-linguistic variation with respect to several phenomena, including modification possibilities. Thus languages such as Finnish and Bemba are claimed to have Verb-selecting causatives because unlike manner adverbs in English, manner adverbs in these languages can modify the caused event, allowing for the low reading to be possible. In Section 2.2, I discuss how the two hypotheses represented in (4) instantiate the learning challenge of the subset scenario that was introduced in Chapter 1. Presently, I will focus on Pylkkänen's account of the ZDC data in English.

The inchoative counterparts of ZDCs are formed simply by merging a verbalizing little v-head to the projection of the root. A partial structure of (3), then, would be as in (5).[3]

(5) *Inchoative* *(Variant I)*

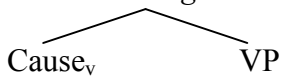
As Pylkkänen does not actually provide an argument that the lexical roots in these causative structures are in fact category neutral, a slight variant of Pylkkänen's analysis would be to assume that the lexical roots have categorical features. According to this variant, √P could be labeled VP. One reason in favor of this alternative proposal is that all the causatives that for Pylkkänen are Root-selecting appear to be roots that occur independently as verbs. This is true for English ZDCs, as well as for what Pylkkänen calls Root-selecting causatives in Japanese. This is a striking distributional fact, which is unexpected (though not incompatible) with the complement of the Cause-head being category-neutral. Although substituting VP for RootP might appear to be a significant departure from Pylkkänen, it is not clear that anything in Pylkkänen's theory crucially hinges on there being a category-neutral root. As far as I can tell, such a change results in the theory maintaining the same empirical coverage, as well as making it more comparable to other syntactic proposals of causatives such as Hale and Keyser (2002) and Ramchand (2008), which assume the Cause-head selects something verbal. As we shall see in Pylkkänen's account of the adverb facts from (1), what is crucial is simply that there is a difference between the lexical root and the functional morphology that

[3] Thus in Pylkkänen's theory, the unaccusative and causative counterparts of these alternating verbs are not derived from each other in the (lexical-)syntax. This contrasts with some other theories of causatives such as Hale and Keyser (2002) and Ramchand (2008). Again, the learning challenge and the model's response to this challenge do not hinge on this. In the discussion of the generalized version of the model in Section 3.2 I comment briefly on how it could be applied to these theories that do derive the causative from the inchoative. However, such a theory would require a different account of the adverbial facts in (1) (cf. note 6), so I will continue to focus on Pylkkänen's theory and close variants of it so as to more closely follow the original learning challenge as it appears in the literature.

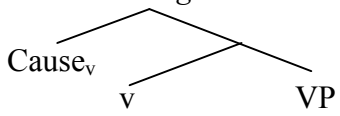
selects it. Moreover, the learning challenge, which is the focus of this chapter is unaffected by this change from category-neutral roots. Under this variant, the structures for the Cause-selection parameter and for inchoatives would be as in (6) and (7) respectively.

(6) *Cause-selection (Variant II)*

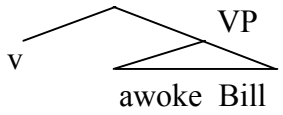
a. *Root-selecting Cause*



b. *Verb-selecting Cause*



(7) *Inchoative (Variant II)*



For the sake of consistency, I will continue to refer to the structures in (4a)/(6a) as Root-selecting, and those in (4b)/(6b) as Verb-selecting.

Assuming that roots have a category, as in (6) and (7), raises a question about the role of the little v-head in (6b) and (7). The role of this head is clear in Pylkkänen's account: it verbalizes the root. However, such verbalization no longer appears to be necessary under this variant. There are various analytical possibilities as to what the role of this head could be. Here I will just mention some but will not develop any further analysis of these ideas.

One possibility regarding the little v-head would be to follow Legate (2003) in assuming that all verbs, including unaccusatives and passives, have a little *v* phase-head as part of their structure (though see den Dikken 2006 for a dissenting view). A variant of this approach is to assume that these verbs have a 'defective' or non-phasal little v-head (perhaps in order to be selected by T). The phasehood of inchoatives would gives us the structure in (7). Once we adopt the structure in (7) for unaccusatives, then the Verb-selecting hypothesis in (6b) is rather straightforward as a hypothesis for the learner. This hypothesis would take causatives to be derived syntactically from unaccusatives (cf. Hale and Keyser 2002; Ramchand 2008). Yet another approach would be to posit some semantic role for the little v-head in (7). Much recent work has looked at decomposing lexical semantics into different 'flavors' of little *v* with different syntactic projections, each with its own semantic contribution (cf. Kratzer 1996; Folli and Harley 2005; Ramchand 2008). A rather natural step is identify the little v-head in (7) with some of the lexical semantics of the root. This step again makes (6b) a viable hypothesis in terms of building the causative from the unaccusative. Any little v-head that is compatible with the meaning of the root would suffice, however if the target grammar for English ZDCs is

(6a), as Pylkkänen proposes, then a potential complication of this semantic approach involves the absence in the structure of (6a) of the little v-head from (7). Given the similarity in the semantics of causative verbs and their inchoative counterparts (cf. Levin and Rappaport-Hovav 1995; Ramchand 2008), one would need to carefully assess what the semantic contribution of this little v-head is. Nevertheless, both the phasehood approach and the lexical semantics approach allow us some basis of understanding the structures in (6) and (7) if we do not adopt category-neutral roots.[4]

Having introduced the Cause-selection parameter, I next introduce the semantic contribution of the Cause-head so as to better understand the modification possibilities in (4a)/(6a) vis-à-vis those of (4b)(6b), and how they relate to the high and low readings for (1). To do this, I briefly mention a second of Pylkkänen's parameters, the Voice-bundling Parameter, which pertains to whether properties of the Voice-head 'bundle' to form a single syntactic head with the Cause-head.

In discussing the Voice-bundling Parameter, an important consideration is that for Pylkkänen, a causative morpheme can be featurally complex. Thus the external argument of a causative (i.e. the *causer*) is introduced by the Voice-feature of the Cause-head. In contrast, the contribution of the Cause-feature of the Cause-head is to introduce a causing event, which is in a CAUSE-relation with the caused event of the lexical root. The meanings of the different components that Pylkkänen proposes for a Cause-head are given in (8). The Voice component in (8a) follows Kratzer's (1996) Voice-head. The Cause component in (8b) takes the event-denoting expression of the caused event, takes a causing event, and says that they are in the CAUSE relationship with each other.

(8) *The Cause-head in zero-derived causatives is:*
a. Voice: $\lambda x.\lambda e.\ \theta_{Ext}(e, x)$
b. Cause: $\lambda P.\lambda e.\ (\exists e')\ P(e') \ \&\ CAUSE(e, e')$

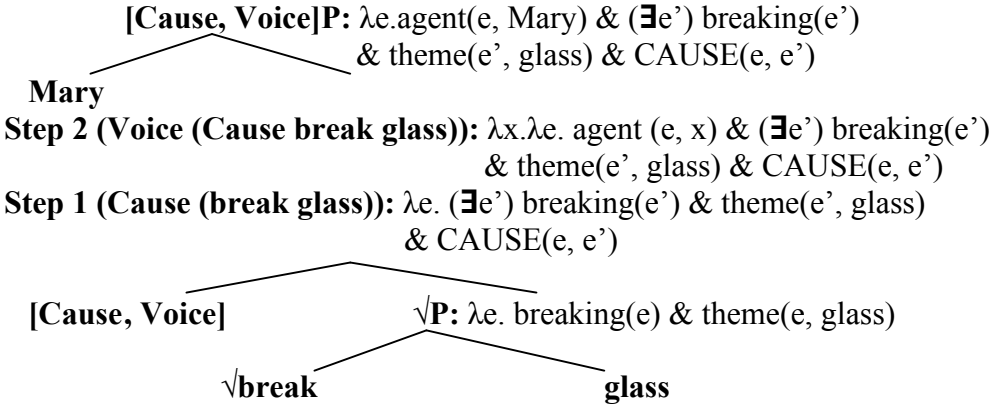
Note that the Cause component of the Cause-head does not introduce the *causer*, but rather the causing event. Still, it is a basic observation that the causative use of verbs such as *break* or *melt* is transitive, i.e. zero-derived causatives always introduce a *causer* that is a semantic individual of type *e*. This is not true of all languages, as Pylkkänen discusses: adversative causatives in Japanese and desiderative causatives in Finnish do not always introduce into the structure an individual-denoting *causer*, although they do introduce a causing event. To account for this cross-linguistic variation – an obligatory *causer* in English and an optional *causer* in Japanese and Finnish – Pylkkänen proposes the Voice-bundling parameter. If a causative morpheme is Voice-bundling, then it must 'bundle' the features of Voice and Cause in a single head in the syntax. In the semantics,

[4] Both these approaches could lead to the learner having additional expectations under the Verb-selecting analysis than under the Root-selecting one. In addition to the absence of the low reading of certain modifiers, then, these could constitute additional sources of implicit negative evidence that the model could learn from. I will not consider these additional sources, but as discussed in Appendix 2, these are certainly welcome sources of evidence for the model, as they can help in expediting the learning process.

the two components of this complex head compose stepwise, with Cause composing first. The *causer*, then, is to be understood as the individual-denoting argument of Voice; and as Cause cannot be introduced without Voice in a bundling language, causative verbs in such a language always occur with a *causer*.

Anticipating the discussion of English being Root-selecting a little bit, Pylkkänen's structure for the bundling causative *break* in English is given in (9) below (Pylkkänen: 99-101). In contrast, if the morpheme is not Voice-bundling, then the Cause-head can simply appear with no Voice component, resulting in an intransitive causative without an overt *causer*.

(9) *Causative 'break' in English*



The semantics of inchoative *break* for Pylkkänen would simply be the sub-tree in (10) whose mother node is √P. Given Pylkkänen's semantics, although the causative is not derived from the inchoative in the syntax, the meaning of the unaccusative structure in (5)/(7) is identical to that of the different complements of the Cause-head in (4)/(6). In this way Pylkkänen can account for the similarity in meaning between the causative and unaccusative uses of the verb.

With the syntactic structures in (4)/(6) and the semantics in (8) in hand, I now return to Pylkkänen's argument as to why the low reading is impossible in (1). Pylkkänen begins by assuming that manner adverbs such as *grumpily* or *happily* are verbal modifiers. That is, they can syntactically attach to verbal projections, but because they are not root modifiers, these manner adverbs cannot attach to the √P. For the variant of Pylkkänen's analysis in (6), we can simply maintain that these adverbs are not root modifiers. Thus they cannot modify VP, but can modify higher verbal projections, or vP shells. What is crucial under both accounts is simply that there is a difference that the modifier is sensitive to.[5]

[5] The discussion above about what type of phrase a modifier can modify raises an additional learning question that I will leave as a matter for further research. The learning question is how the learner knows that some modifiers, such as manner adverbs, do not modify roots. Pylkkänen suggests that evidence these

Under either analysis of the inchoatives in (5) or (7), a modifier of vP shells can adjoin to vP. As we saw in (3), repeated below, this results in modification of the event whose participant is the internal argument of the unaccusative verb. The modification structure is given in (10).

(3) Bill awoke grumpily/happily.

(10) [Bill … [$_{vP}$ [$_{vP}$ *v* [awoke ~~Bill~~]] Adv]]

I next consider modification possibilities for causatives. Under both formulations of the Cause-selection parameter in (4) or (6), with a Root-selecting causative, there is only one verbal attachment site, namely adjoining to CauseP in (4a)/(6a). In contrast, a Verb-selecting causative provides two verbal attachment sites in (4b)/(6b): adjunction to the vP projection of the little v-head and adjunction to the CauseP. The fact that Verb-selecting cause provides more options for adjunction corresponds to a difference in interpretive

---

adverbs cannot modify roots comes from the impossibility of **a grumpily awake boy*. Pylkkänen's argument is that the impossibility of modifying the adjectival use of *awake* with grumpily indicates that *grumpily* cannot adjoin to any projection including the projection of the lexical root. This presupposes that there are no little v-heads in the structure of these adjectives; as a verbal modifier, the adverb would presumably be able to adjoin to such a vP, contrary to what we see in Pylkkänen's example. It should be noted, though, that Pylkkänen gives no other evidence that these adjectives do not contain a vP layer in their structure. However, although not particularly common, examples of this kind of modification are attested, as in (i):

(i) 6:25 came awfully soon, but at least it came with the sounds of a happily awake baby and the sight of a gummy grin.
(Accessed 21 April 2015: http://amysfinerthings.com/a-day-in-the-life-january)

The possibility of modification in (i) is not particularly informative as to the place of adjunction. For example, the adverb could adjoining to a little a-head, or even a little v-head if vPs are possible in the structure of these adjectives.

A perhaps more promising possibility is that the set of root-modifiers falls out from some relatively invariant cross-linguistic hierarchy of adverbs in the spirit of work such as Cinque (1999). It is worth noting that the types of root-modifiers identified by Vecchiato (2011) are largely what Cinque refers to as 'circumstantial' modifiers. The circumstantial modifiers are located at the bottom of Cinque's hierarchy, but Cinque claims that these modifiers are not ordered with respect to each other. More recent work, such as Schweikert (2005) and Cinque (2006), argues that there is some ordering among these modifiers. However, these works do not consider where manner adverbs might be ordered. Vecchiato considers the possibility that there might be a correlation between root-modifiers being lower on the hierarchy and non-root-modifiers being higher on the hierarchy. Ultimately Vecchiato rejects this approach, though the reason for doing so is not clear. Vecchiato notes that manner adverbs, which do not modify the root, are low on the hierarchy and are among the group of circumstantial modifiers. As mentioned above, though, developments on the ordering of circumstantial modifiers have not looked at manner adverbs, leaving open the possibility that manner adverbs are ordered higher than other circumstantial adverbs. Vecchiato also notes that degree modifiers (e.g. *completely*), which are root-modifiers, are low on the hierarchy but are higher than circumstantial modifiers. Given the connection between degree modifiers and the telicity/end-state of events that a number of researchers have drawn (e.g. Tenny 2000; cf. also Cinque 1999; Ernst 2002), there might be independent semantic reasons underlying degree modification of the root, which encodes the end-state semantics in (9). In sum, the research program outlined in Cinque's work offers an area of potentially fruitful research on this topic.

possibilities for the two structures. Recall from the semantics in (8) that the Cause-head introduces a causing event, at which point in the derivation the caused event is existentially closed. Pylkkänen's argument is based on the following assumption about how event semantics are computed: when the lower caused event is existentially closed in (8), event modifiers such as *grumpily* and *happily* can modify only the higher causing event introduced by the Cause-head. Thus lower event modification for verbal modifiers is simply impossible in (4a)/(6a), and this is an immediate consequence of the structure, given that there are no verbal projections below the Cause-head. In the structure for Verb-selecting cause in (4b)/(6b), though, modification of the lower caused event is possible just in case the verbal modifier adjoins to the lower vP projection. The only way for the low reading to be possible, then, involves adjunction to vP in (4b)/(6b). Why, then, does such adjunction appear to be impossible? Indeed, such adjunction is not ruled out by Pylkkänen's semantics (nor is it in the semantic representations of Levin and Rappaport-Hovav 1995 or Ramchand 2008). Given that the vP in (4b)/(6b) has the same meaning as the unaccusative (cf. the discussion above), and given that modification of the unaccusative is possible as in (3), we have no semantic grounds in this theory for blocking the low reading in the causative. Pylkkänen's answer is to exploit the structural possibilities of complement size made available in the Cause-selection parameter. Given that the low reading is not available in the causative in (1), Pylkkänen concludes that there must be no vP projection in the structure of the causatives, a criterion that can be satisfied only with Root-selecting cause. Thus the simpler syntactic structure of Root-selecting cause derives the lack of ambiguity with a verbal modifier in (1).[6]

In contrast to manner adverbs modifying English ZDCs, manner adverbs in Finnish and Bemba do allow for modification of the caused event, giving rise to the low reading. Assuming that manner adverbs are cross-linguistically not root modifiers, Pylkkänen takes such low readings to be evidence that these languages have a Verb-selecting Cause-head (see Section 3.4 for additional evidence that these languages are Verb-selecting). It is only with a v-head intervening between the root and the Cause-head that these adverbs could modify the lower sub-event.

Turning now to a learning perspective of Pylkkänen's argument, the adult grammar, which allows only the high reading in (1), can be taken to be the target state for the learner's grammar; this target state will be taken as evidence that the learner has the correct parameter setting. What we see in (4)/(6) is that the syntax of a Verb-selecting cause is structurally more complex than that of a Root-selecting cause. Pylkkänen's claim is that examples such as (1) show that the zero-derived causatives in English are Root-

[6] Theories such as Hale and Keyser (2002) and Ramchand (2008) would need a different account for the absence of the low reading in (1). In these theories, an inchoative structure such as (5)/(7) is the complement of the causativizing head. If a manner adverb can modify the unaccusative as in (3), then the question is why this would not be possible in the causative in (1). As we have seen, a theory such as Pylkkänen's in which the Cause-head can take a smaller complement, provides a straightforward account of the data.

selecting and thus instantiate the simpler structure. Assuming Pylkkänen is correct, the central question of this chapter concerns learning a parameter setting of Root-selecting (4a)/(6a) over that of Verb-selecting (4b)/(6b) for these causatives in English.

Focusing on the learnability issue related to the Cause-selection parameter, I will not discuss the Voice-bundling Parameter further. To simplify subsequent discussion of learning the Cause-selection parameter, I will simply assume that the Voice-bundling parameter has already been learned correctly. As far as I can tell, the Cause-selection and Voice-bundling parameters can be learned independently, and this simplification does not alter in any substantive way the learnability question with respect to the Cause-selection parameter. I now turn to the challenge that learning the Cause-selection parameter raises.

2.2 The learning challenge

I will assume that Pylkkänen's argument about zero-derived causatives in English being Root-selecting is correct. The question I will pursue is whether these causatives can correctly be learned as being Root-selecting given the evidence that Pylkkänen discusses. At first blush this does not appear to be possible. The reason for this is because from the perspective of the learner, the data in (1), repeated in (11) below, underdetermine which analysis (Root or Verb-selecting) is correct.

(11) John awoke Billy grumpily/happily.

In order for a grammar to account for (11), it must be able to generate a string-meaning pair in which (among other things) (a) a Cause-head embeds a root, and (b) the modifier adjoins to CauseP, thereby modifying only the causing event. Now, a grammar with a parameter setting of either Root-selecting or Verb-selecting cause is able to generate such output, as is clear from the preceding discussion. Given this indeterminancy, either hypothesis would appear to be a viable option for the learner.

In fact, the learning challenge is more general. Note that the same indeterminancy is true for the non-modified example in (12).

(12) John awoke Bill.

To generate the example in (12), the grammar does not even need to consider the issue of which projection an adverb is adjoining to and which event it is modifying – the two parameter settings are seemingly equally good at providing Cause-heads that embed lexical roots. In Section 3.2, I show how the learning model can exploit the more general nature of the learning challenge so as to be able to learn the Root-selecting setting on the basis of input such as (12). As will be discussed further, this is desirable to the extent that the input in (11) is infrequently attested in the learner's input.

Recall that Pylkkänen's argument crucially involved considering the impossibility of the low reading (i.e. negative data), a reading that a child will presumably never be

exposed to in the primary linguistic data. Given that there is no clear positive evidence in favor of the Root-selecting hypothesis, we are left with the following acquisition challenge: how do children correctly choose between Root-selection and Verb-selection for Cause-selection? Pylkkänen's argument relies on negative evidence, but how can children learn from this evidence?[7]

Before discussing how the learning model can address this challenge (by crucially capitalizing on the fact that a learner never hears the low reading), I frame the learning challenge in the context of the subset scenario that was introduced in Chapter 1. To begin with, if we consider the structural and interpretive properties of the two causative structures in (4)/(6), we see that those of Root-selecting cause are a proper subset of those of Verb-selecting cause. Thus in (4), for example, the following holds: (a) the core set of syntactic heads is {Cause_v, √} for Root-selecting and {Cause_v, v, √} for Verb-selecting; (b) the set of verbal adjunction positions is {CauseP} for Root-selecting and {CauseP, vP} for Verb-selecting; and (c) the set of interpretive possibilities for verbal modifiers is {high reading} for Root-selecting and {high reading, low reading} for Verb-selecting. Given Pylkkänen's semantics in Section 2.1, and holding all other parameter values constant, then crucially the set of utterances (string-meaning pairs) that can be generated by the Root-selecting grammar are a proper subset of those generated by the Verb-selecting grammar. The point of difference is that the Verb-selecting grammar can generate strings with the low reading, whereas the Root-selecting grammar cannot. Thus the Root-selecting grammar generates a language that is a proper subset of the language generated by the Verb-selecting grammar. Further, we see that the ZDCs in (11) and (12) instantiate the schematic example of the subset scenario from Chapter 1. The schema from Chapter 1 is repeated below.

(13) a. $[_{XP}\ X\ [_{YP}\ Y\ \ldots\ ]\ ]$
b. $[_{XP}\ X\ [_{ZP}\ Z\ [_{YP}\ Y\ \ldots\ ]\ ]\ ]$

The grammar of the subset language in (13a) and the grammar of the superset language in (13b) correspond to Root-selecting and Verb-selecting causatives respectively. X represents the Cause-head, YP the projection of the lexical root, and Z is a phonologically null little v-head. As we have seen, both structures are compatible with all the learner's input, underdetermining the correct structural analysis. One way to address this learning challenge, which I will not adopt, would be to suppose that the learner can choose the grammar of the subset language on the basis of the 'Subset Principle' (Berwick 1986; cf. Wexler and Manzini 1987).

[7] The learning challenge can in principle be recreated for theories such as Hale and Keyser (2002) and Ramchand (2008), in which the complement of the Cause-head is a vP that embeds the root. All that is necessary is a hypothesis space that considers an alternative structure in which there are two vP shells embedding the root. If both hypotheses are compatible with the input, then we are left with a similar learning question regarding which structure the learner will adopt. See also note 9 for discussion of these theories in light of the generalized implementation of the learning model in Section 3.2.

One way to state the Subset Principle would be the following: given two hypotheses A and B such that A can be considered a narrower hypothesis that B (in terms of a proper-subset relation), do not consider B unless forced to do so by the input. If we Root-selecting cause to instantiate a narrower hypothesis than Verb-selecting caseu, as per the discussion of subset/superset languages above, and given that both structures adequately account for the modified and non-modified data in (11) and (12), one could invoke the Subset Principle in the following way. Children learning zero-derived causatives in English only ever consider the simpler Root-selecting structure, and are never forced to consider the more complex Verb-selecting structure (because, for example, they never hear such a causative with a low reading, which cannot be generated with the Root-selecting structure).

A similar point also holds for a default parameter setting, which is another possible way of addressing the learning challenge of the subset scenario. One could suppose that children have a default parameter setting of Root-selecting that only switched to Verb-selecting given the appropriate triggering input (such as low adverbial modification).

The contribution of the learning procedure I propose is that the simpler or 'subset structure' can be learned without needing to invoke either a principle that achieves this result or a default parameter setting. Thus the learner can consider both superset and subset hypotheses during the course of learning, but the actual process of learning will lead the learner to the simpler hypothesis. I now turn to the learning procedure with respect to zero-derived causatives in English.

3. Addressing the challenge

The question Pylkkänen's argument raises is whether or not children can reliably learn the target grammar given the input data that is available to them. This grammar is the grammar of the subset language. In this section, I propose that children can make use of implicit negative evidence to learn the correct parameter setting. The proposal is that the learning procedure is sensitive to the *absence* of evidence that would unambiguously determine the more complex structure as correct. In the model, under the more complex structure of the grammar of the superset language, the learner has an expectation (or probability) that such evidence will occur. Given that such evidence never occurs, the learning process will ultimately settle on the simpler structure, for which there is no such expectation. This section presents two implementations of the model that formalize this proposal. These implementations make use of the generative capacity of the full version of the model introduced in Chapter 2. Although the model has various simplifying assumptions regarding the acquisition of English, it provides a clear way of showing that on a basic and abstract level, such a model is up to the learning challenge: despite data that apparently underdetermine the correct structural analysis, the model succeeds in learning correct parameter setting from implicit negative evidence.

Two implementations of the model are developed here. In the first version, presented in Section 3.1, I focus on the adverbially modified data that underlie Pylkkänen's (2008) account, and discuss how the model can learn the correct parameter setting from this input. I then go on in Section 3.2 to present a minimally different implementation of the model, which shows how the learning process is more general and is applicable to even non-modified data. Thus correct parameter setting can be learned from run-of-the-mill non-modified data, a desirable consequence given the relative rarity of the modified data. Results of running this second implementation of the model are given in Section 3.3. I conclude in Section 3.4 with further discussion of what kind of evidence can allow the model to learn the grammar of superset language. Languages with the parameter set to the more complex structure do of course exist, such as Finnish, but the prediction here is that they must be different from English with respect to what kinds of evidence lead to parameter setting for the more complex structure.

3.1 Learning from implicit negative evidence: The case of zero-derived causatives

In this section, I outline the basic proposal for how the model could learn a simpler structure (Root-selecting cause) given evidence that underdetermines the correct structural analysis. The model presented in this section directly concerns the kind of modified data that Pylkkänen discusses. The insight of the proposal is that children learn via implicit negative evidence. In this framework, each analysis (Root or Verb-selecting) is associated with a prior probability that represents the learner's expectations regarding the shape of the adult grammar. Further, the more complex Verb-selecting analysis comes with the expectation that its greater range of structural and interpretative possibilities will be encountered in the primary linguistic data. This expectation can also be expressed with a probability. However, this expectation is not met, and it is this absence of the evidence for the low adverbial reading that the model is sensitive to. Over time, as the expectation continues to not be met, the probability of the analysis given the input will decrease to the point where it can be discarded by the learner as a plausible analysis. What remains is the simpler Root-selecting hypothesis, whose likelihood given the input has been increasing while that of the alternative Verb-selecting hypothesis has been decreasing.

The implementation here involves the full version of the model, as introduced in Chapter 2. Recall that this is a generative model. That is, the model goes through a series of probabilistic choices (one of which is the Cause-selection parameter) to generate output of a string-meaning pair. Further, the model can be organized so that some choices are conditional on having made other choices. The model then checks to see whether this output matches the input. If there is a match, then the model will reinforce the choices used to generate that output by increasing the weights associated with those choices. By doing so, the model increases the likelihood of making those successful choices again when presented with similar input. Thus the implementation of the model here goes beyond simply checking to whether competing grammars are compatible with the input,

as was the case in Chapters 3 and 4. To determine a grammar of best fit in the cases of Swiss German and Korean, it was sufficient to simply check compatibility because none of the grammars under consideration generated languages that were in a subset/superset relation with each other. This is not the case in the subset scenario with zero-derived causatives. Further, with zero-derived causatives there is only a single parameter under consideration, meaning that parameter interaction does not play a role. Instead, the grammar of best fit here can be determined by measuring which grammar is more likely to generate output that matches the input, or target output. As we shall see, this grammar of best fit is the grammar of the subset language. The focus of this section, then, is on what kinds of output the different hypotheses can generate, and how they do so.

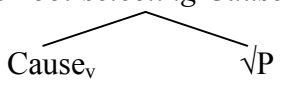
I begin by presenting the crucial choices in the model that are conditionally related to the choice of a value for the Cause-selection parameter, and that are most relevant for an example such as (1), repeated here.

(1) John awoke Bill grumpily/happily.

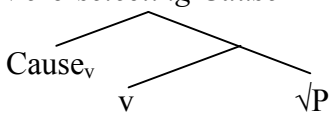
Recall from (4), repeated below, that a Root-selecting structure has only one possible site for adjunction of non-root modifying grumpily-type adverbs (i.e. an adverb that clearly illustrates only the high reading in (1)): the CauseP. In contrast, the Verb-selecting structure allows for two verbal adjunction sites: CauseP and vP. I use Pylkkänen's formulation of the parameter here for illustrative purposes; nothing essential changes should one adopt the variant in (6).

(5) *Cause-selection*

a. *Root-selecting Cause*



b. *Verb-selecting Cause*

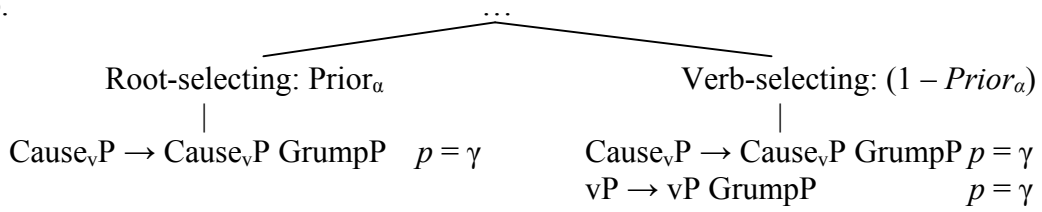


These different adjunction possibilities can be represented as conditional choices in the model, as illustrated schematically in (14). Choices regarding modification are conditional on having first chosen a value for the Cause-selection parameter. In attempting to generate output that matches the input in (14a), the model will first choose a parameter value – either Root-selecting or Verb-selecting – which is determined probabilistically by Prior_{α} / (1 – Prior_{α}). The model will then choose whether there is grumpily-type modification and where it adjoins, again probabilistically. Thus choosing to modify vP, for example, is conditional on having chosen Verb-selecting. Only if the model exclusively chooses to have this modification adjoin to CauseP can it generate the correct output. Further, I will assume that there is an expectation for grumpily-type modification with a probability $p > 0$, and that p is the same regardless of the vP shell being adjoined to. I will use the variable γ to refer to this probability, and discuss what it

might be below. This is shown in the partial schema below, which gives a subset of the choices in the model for both Root and Verb-selecting grammars. I use GrumpP as a catchall category for any sort of grumpily-type modifier in the phrase structure rules in (14b).

(14) a. Input: *John awoke Bill grumpily.* (high reading)

b.



The question of what the probability of a verbal projection being modified by a grumpily-type adverb is is somewhat less straightforward. One way of approximating this probability would be to calculate it via a frequency rate of sampled vP shells from a corpus. I will not attempt such a task here. Indeed, this probability could also be generated from a dirichlet distribution that is updated during learning to reflect the frequency of the modifiers in the input. This would be in line with how I have been modeling the learner's expectations regarding parameter values throughout this thesis (cf. below for a similar treatment of the Cause-selection parameter itself). Instead, for practical purposes I will plug in a variety of different values for this probability for different runs of the model; for each run that value will remain constant. This kind of flexibility allows the model to easily accommodate whatever the results of a raw frequency rate from a corpus search might be. Additionally, this flexibility in what the probability of modification is allows for the possibility of clearly testing what the limits of the model are. As we shall see, the larger the probability for γ, the more likely it is that the Root-selecting grammar can be learned as the grammar of best fit. The test would be to see how low the probability can go before the model is no longer able to consistently learn Root-selecting parameter setting. These are not questions that I will address in this chapter; rather they are questions that have influenced the design of the model with the intent of expanding its empirical accuracy.

Thus we see that the range of expectations in grumpily-type adverbial modification varies depending on the choice of Cause-head. Under a Root-selecting cause-head there is an expectation only for modification that would result in a high reading, whereas a Verb-selecting cause-head has expectations for modification that results in either a high or low reading. This extra expectation is the implicit negative evidence that the model is sensitive to, and as this extra expectation is never met (i.e. there is never any input instantiating the low reading), it is this extra probability (for a low reading) that ultimately causes a Verb-selecting grammar to be less likely. To see how the Verb-selecting grammar is less likely, I introduce a fuller illustration of the choice-points of the model below, which will be able to generate output like (14a).

Before presenting a fuller picture of the choice-points of the model, I note some simplifications in the model. First, the corpus that the model learns from will be simplified to contain only utterances of the form in (14a). And I will assume that children correctly understand these tokens as having the target interpretation of the high reading. (Also, as mentioned in Chapter 2, I assume an idealized leaner with adult-like compositional semantics, such that the high reading results from adjunction to CauseP.) As an empirical matter, though, to the extent that children struggle to interpret this input correctly, it provides further support to base learning off non-modified examples (as in the revised model in Section 3.2), which have no modifiers whose interpretive properties could be misconstrued. Such a simplified corpus results in a learning task that is admittedly far simpler than what we expect a child to actually experience. However, the model is still able to distill what is essential in the learning challenge for zero-derived causatives.

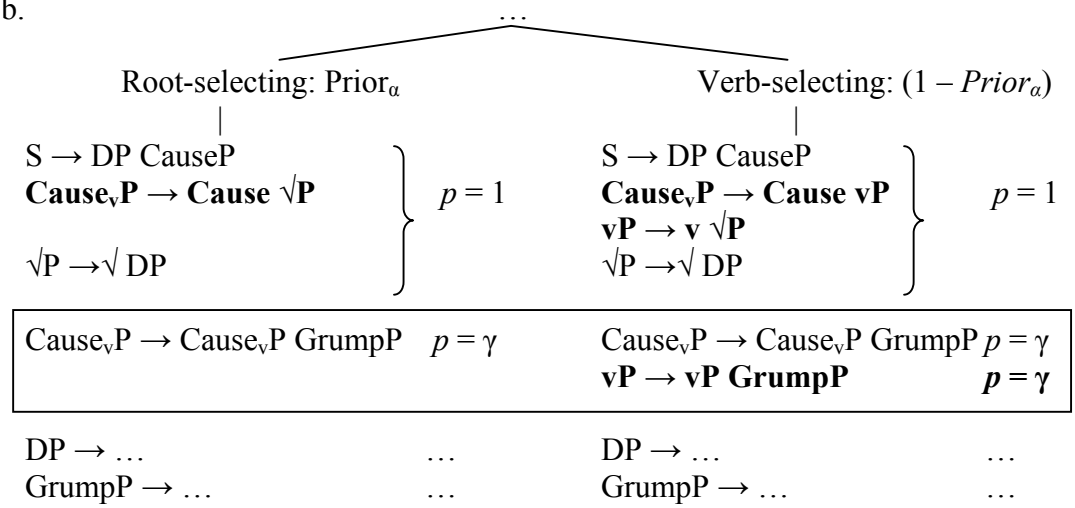
Second, the additional phrase structure rules presented below and their accompanying probabilities have been, in a sense, reverse-engineered to streamline the learning process here. As we saw in Chapters 3 and 4 regarding word order in Swiss German and Korean, the model is also conditioned to learn phrase structure relating to word order (e.g. head-complement order), and during the learning process the model can output strings that look decidedly non-English. In order to focus in particular on learning the Cause-selection Parameter, the model will reflect what we already know about English phrase structure, as if the model had already been trained and conditioned with respect to these properties so as to more expediently generate the attested forms. Thus the probabilities given below will all be 1, except for those of the crucial choices already given in (14) that distinguish the two grammars. This allows us to abstract away from any choices that, by hypothesis, will be identical across the two Cause-selection grammars. It also reflects that there is in no variation in the simplified corpus. In the spirit of this simplicity, I again abstract away from additional functional projections such as CP and TP (hence a rule such as S → DP CauseP), and do not fully expand some phrasal nodes (e.g. DP), or include terminal nodes (e.g. *Bill*). Again, the only probabilistic difference between the two grammars concerns the choices in (14); everything else is, by hypothesis, equal, and thus I abstract away from it. A line of future research is to expand the model so that it learns from a corpus that better reflects the primary linguistic data. Nevertheless, such additional complexity is not predicted to affect learning Cause-selection parameter setting in any substantive way.

A schema of the choice-points in the model is given in (15) below. Differences between the two grammars are given in boldface, and the crucial different expectations with respect to modification are boxed. Note that the prior probability for the choice of Root-selecting or Verb-selecting grammar has been left rather vague. As with other parameters in previous chapters, this probability will be drawn from a dirichlet

distribution, which is discussed in Section 3.3. I also assume that the number of adverbs modifying the same phrase is limited to one.

(15) a. *John awoke Bill grumpily.* (high reading)

b.



(15) shows all the relevant choices that the model needs to make to be able to generate the attested input. Beginning with a choice of Root or Verb-selecting parameter value, the model then makes choices with respect to grumpily-type modification and generates the rest of the phrase structure.

We are now in a position to see how the extra expectation of low modification can lower the overall likelihood of the model generating output with a Verb-selecting structure. Given the input in (15a) with the high reading, consider what the probabilities – $p(G_{Root})$ and $p(G_{Verb})$ – are of generating output that matches that input under the two grammars. To calculate this for the Root-selecting grammar, because all other probabilities are 1, we simply need the joint probability of choosing the Root-selecting value (Prior$_\alpha$) and the probability of GrumpP adjoining to CauseP (γ). For the Verb-selecting grammar, this is the joint probability of the choosing the Verb-selecting value (1 – *Prior*$_\alpha$), the probability of GrumpP adjoining to CauseP (γ), and the probability of GrumpP *not* adjoining to vP (1– γ):

(16) *Probabilities of different grammars generating attested input* [John awoke Bill grumpily.] (high reading)

a. $p(G_{Root}) = p(CauseP\ GrumpP|CauseP)p([Cause\ \sqrt{P}]|CauseP) = (\gamma)(Prior_\alpha)$

b. $p(G_{Verb}) = p(CauseP\ GrumpP|CauseP)p([Cause\ vP]|CauseP)\ p(\neg vP\ GrumpP|vP) = (\gamma)(1 - Prior_\alpha)(1-\gamma)$

As in Chapters 3 and 4, the priors for the Cause-selection parameter (Prior_α / $(1 - \textit{Prior}_\alpha)$ above) will be drawn from a dirichlet distribution (see Section 3.3 and Chapter 2 for more details). For the sake of illustration here, though, let us simplistically assume that there is a 50-50 chance of choosing either grammar. Substituting in .5 for Prior_α and $(1 - \textit{Prior}_\alpha)$, what we see is that regardless of the value of γ (the probability of a vP shell being modified by a grumpily-type adverb), the probability of generating the attested data is less under the Verb-selecting grammar than for the Root-selecting one:

(17) *Probabilities of different grammars generating attested input* [John awoke Bill grumpily.] (high reading) *with priors of .5*

a. $p(G_{Root}) = (\gamma)(\textit{Prior}_\alpha) = (\gamma)(.5)$

b. $p(G_{Verb}) = (\gamma)(1 - \textit{Prior}_\alpha)(1 - \gamma) = (\gamma)(.5)(1 - \gamma)$

In this way, then, the absence of expected evidence lowers the overall likelihood of the more complex grammar. So long as the value of γ (i.e. the probability of modification of a vP shell) is sufficiently greater than 0, the difference between (17a, b) will favor the Root-selecting grammar, thereby pushing the learner toward adopting the parameter setting of the simpler grammar. This means that the learner is more likely to reinforce the Root-selecting parameter setting. With exposure to more input, the learner will on average reinforce the Root-selecting value more and more as they are gradually pushed toward learning the grammar of the subset language.

I have now presented the basic outline of how the generative model can address the learning challenge of ZDCs in English. In the results of Section 3.3 we will see that under the revised implementation of the model, the value for γ can go rather low with the model still having success in learning the grammar of the subset language. I next turn the to revised implementation of the model.

3.2 Making the model more general

In the previous section I outlined the basic proposal for how implicit negative evidence can be informative for the learner, allowing for the correct parameter setting. I have followed Pylkkänen in assuming that the parameter setting depends on the absence of the low adverbial reading. However, one might wonder whether a child actually learns the correct parameter setting given the apparent rarity of utterances like the example in (1), repeated below, utterances that have both a zero-derived causative and the appropriate verbal modifier that clearly exemplifies only a high reading and not a low reading.

(1) John awoke Billy grumpily/happily.

For instance, a simple search in The Corpus of Contemporary American English (Davies 2008) for these kinds of adverbs co-occurring with common causatives such as *open*, *close*, and *move* resulted in only a handful of examples, as compared to thousands of examples of non-modified transitive uses of these verbs. This is admittedly only an impression of the relative frequencies of these kinds of data, but they are indicative of a potential complication for the learner if the learner could set the parameter only by paying attention to the relatively infrequent data. Nevertheless, an advantage of the proposal above for learning from implicit negative evidence is that it is sufficiently general to extend to the fairly common case of non-modified causatives, such as the example in (12), repeated below.

(12) John awoke Bill.

In fact, only a few changes need to be made to the model in Section 3.1 above. Thus the rate of grumpily-type modification (however small) is not actually critical, and the kinds of examples carefully constructed by linguists such as (11) are simply an extension of a more general part of the model. In the remainder of this section I briefly show how the model can be changed to learn parameter setting from non-modified examples such (12). I report results of simulations of this more generalized version of the model in Section 3.3.

To extend the proposal above to the more general case of the non-modified ZDC in (12), we only need to consider the probabilities of *any* modifier of a vP shell under the two hypotheses about causative structure. Thus instead of considering only the probability of, say, a grumpily-type verbal modifier as in (11), we now consider the whole host of adverbials that modify vP shells and that could appear with a given causative structure. The idea here is that the absence of any modifier is a kind of implicit negative evidence. All things being equal, the bigger the structure, the more adjunction positions there are, and the more likely it is for a modifier to appear. This is shown schematically in (18), in which the likelihoods of different output are a direct reflection of the learner's expectations of the shape of the input: modification is expected more under a Verb-selecting grammar than under a Root-selecting one.

(18) *Expectations of modification*

$$\underbrace{p(\textit{Modified-CauseP}) \vee p(\textit{Modified-vP})}_{\text{Verb-selecting}} \quad > \quad \underbrace{p(\textit{Modified-CauseP})}_{\text{Root-selecting}}$$

Conversely, given that there are no modifiers in the input, a simpler structure (under which modification is less likely) is more compatible with the input. The absence of

modification is expected more under a Root-selecting grammar than under a Verb-selecting one:[8]

(19) *Expectations of <u>no</u> modification*

$p(\neg Modified\text{-}CauseP) > p(\neg Modified\text{-}CauseP) \wedge p(\neg Modified\text{-}vP)$

Root-selecting Verb-selecting

Again, I assume that the probability of modifying any vP shell is equal across vP shells, and that $p > 0$. As I will run the model with various values for this probability, as described in Section 3.1, I keep it as the variable γ. The schema of choices in (20) can be compared with (15), where GrumpP has simply been replaced by AdvP, a catchall category for any sort of relevant adverbial modifier.

(20) a. *John awoke Bill.*

b.

…

Root-selecting: $Prior_\alpha$		Verb-selecting: $(1 - Prior_\alpha)$	
S → DP CauseP	$p = 1$	S → DP CauseP	$p = 1$
$Cause_vP$ → Cause √P		**$Cause_vP$ → Cause vP**	
		vP → v √P	
√P →√ DP		√P →√ DP	
$Cause_vP$ → $Cause_vP$ AdvP	$p = \gamma$	$Cause_vP$ → $Cause_vP$ AdvP	$p = \gamma$
		vP → vP AdvP	$\boldsymbol{p = \gamma}$
DP → …	…	DP → …	…
AdvP → …	…	AdvP → …	…

What we can see from (20) is that the Verb-selecting structure is more likely to have some sort of modification of a vP: the probability of either CauseP or vP being modified is greater than that of just CauseP being modified. In other words, there is more of an expectation to see modification of a vP shell under the more complex hypothesis regardless of what kind of modification this is (PP modification, grumpily-type modification, etc.). This means that input like (20a) without any modification at all has a higher probability of being generated by the simpler Root-selecting grammar where there are fewer expected possible positions for modification. This is illustrated abstractly in (21), which can be compared with (16) and (17).

[8] Thus input with any modifier is actually more likely under the Verb-selecting hypothesis, a point I return to in Section 3.4.

(21) *Probabilities of different grammars generating attested <u>non-modified</u> input* [John awoke Bill.] (high reading) *with priors of* .5

a. $p(G_{Root}) = p(\neg CauseP\ AdvP|CauseP)p([Cause\ \sqrt{P}]|CauseP) = (1-\gamma)(Prior_{\alpha}) = (1-\gamma)(.5)$

c. $p(G_{Verb}) = p(\neg CauseP\ AdvP\ |CauseP)p([Cause\ vP]|CauseP)\ p(\neg vP\ GrumpP|vP) = (1-\gamma)(1-Prior_{\alpha})(1-\gamma) = (1-\gamma)^2(.5)$

The probability of generating non-modified output under the Root-selecting grammar is the joint probability of choosing the Root-selecting grammar and choosing no adverbial modification at the CauseP phrase marker. In contrast, the probability of generating non-modified output under the Verb-selecting grammar is the joint probability of choosing the Verb-selecting grammar and choosing no adverbial modification at both the CauseP and vP levels. If we again use the illustrative prior of .5 for choice of Cause-selection parameter value, the schema in (21) illustrates how the extra expectation of modification adjoining to vP lowers the probability of generating the attested input with the Verb-selecting grammar. So long as the probability of modification is within the interval [0, 1] this difference in likelihood holds. Further, the greater the probability, the less likely the Verb-selecting grammar is to generate target output than the Root-selecting grammar is. As discussed in Section 3.1, with repeated input to the learner, the difference in probabilities will push the learner more and more toward a Root-selecting grammar. On average the learner will repeatedly reinforce the Root-selecting parameter value more than the Verb-selecting one. The Verb-selecting grammar will then be sampled less and less, with the learner eventually settling on the Root-selecting grammar, which is the grammar of the subset language.[9]

Thus even if the child is exposed to non-modified examples of zero-derived causatives (and simple searches for common causatives in CHILDES (MacWhinney 2000) show that there are many such examples), the child can still learn from implicit negative evidence (concerning the absence of modification) that the simpler hypothesis has a higher likelihood of being correct given the input. As the results in the following section show, even with a relatively low probability for modification, the Verb-selecting hypothesis will have leaked probability sufficiently such that the model can be said to have rejected it as a viable parameter setting. The grammar of the superset language, with

[9] The logic behind (21) can be extended straightforwardly to theories such as Ramchand (2008), in which the Cause-head embeds a vP layer. As mentioned in note 7, in this framework an alternative analysis that the learner could consider would involve the Cause-head embedding two (or more) vP shells. However, any structure that is more complex (i.e. embeds more than a single vP layer) will leak probability along the lines of (21), and will be less likely to generate target output than the structure that embeds a single vP layer. The more vP layers, the more likely the output is to have a modifier, and having such a modifier does not match the input. Even in this more complex analysis of causatives, it is the structurally simpler hypothesis that emerges as the grammar of best fit given the input.

its greater number of adjunction sites, is thus not as good a fit to the input as the grammar of the subset language. By using this measure of fitness, the model can address the learning challenge and adopt the grammar with the simpler structure.

3.3 Results

In this section I report results of running a simulation of learning the Cause-selection parameter. The model is run with a program written in the Church programming language (Goodman et al. 2008). I used the *bher* implementation of Church.

The specifications of the model are largely the same as those used for the simulation of Korean in Chapter 4. As the null hypothesis, I assumed all parameter values are initially weighted equally and thus used equal priors. I also used priors with weak expectations for the Cause-selection parameter. Accordingly, the parameter values had weak (or low) initial weights. As all the input to the model is ambiguous, it is important not to have the initial weights be too great for parameter values. With weaker initial weights, it will be easier for the learner to converge on one of the two parameter values for the Cause-head. The Cause-head selection parameter is thus represented with a dirichlet distribution that has prior pseudo-count values of 1 for all parameter values. The pseudo-counts are the parameter weights that represent the learner's expectations regarding the shape of the adult grammar. Each parameter's dirichlet distribution uses its weights to generate a probability for a given parameter value. These probabilities are then used to sample a grammar and generate output for a given token of input. Thus, initially the learner has a dirichlet distribution of *dir*(1, 1), where the first pseudo-count total corresponds to Verb-selecting cause, and the second to Root-selecting cause. For more discussion of the dirichlet distribution, I refer the reader to Section 2.4 in Chapter 2.

The update procedure is also similar to what was used in Chapter 4. For each token of input, the model will generate output that matches the input 100 times. Each time the model generates such target output can be called a chew. I note that in Chapter 4, the model was specified to have only 10 chews per token of input. The increase in the number of chews will be discussed shortly below. These 100 chews represent what the model learns for each token of input. This can be expressed in terms of pseudo-counts. If only one grammar G_1 is selected in all 100 chews, the pseudo-counts of the parameter values that comprise that grammar will be increased by 1. If G_1 is Root-selecting after the first token of input, then the updated pseudo-counts will be the dirichlet distribution of *dir*(1, 2). This means that the adjustment to the weight of any parameter value being used after a single chew is 0.01. If after the first token of input some target output was generated 70 times under Root-selecting cause, while being generated 30 times under Verb-selecting cause, then the updated dirichlet distribution for Cause-selection would be *dir*(1.3, 1.7). This process iterates after each subsequent token of input, with pseudo-count values increasing accordingly. However, to expediently see the proof-of-concept model shift most of the probability mass onto one parameter value over the other (and

thus learn the parameter setting with a heavy weighting) without the pseudo-count values becoming very large, I have chosen to normalize the pseudo-count total of each parameter after a certain amount of input. Again, to model learners that are not too conservative, I normalized the sum of the pseudo-count values to 10 after every 20 tokens of input. This means that the sum of the two pseudo-count values would equal 10 after 40 tokens of input. I note that the normalization process can sometimes result in the weights becoming very small. The pseudo-count values for the dirichlet distribution must be greater than 0, but if the value becomes too low, the model will treat it as if it is 0, resulting in a domain error. To avoid this, a parameter's weights were adjusted to 0.02 and 9.98 if after being normalized, they were below 0.02 or greater than 9.98.

The difference between the learning procedure in Chapter 4 and the one adopted here concerns the number of chews. In Chapter 4 there were only 10 chews per token of input, whereas here there are 100. As discussed in Chapter 2, such an increase represents a more tentative learner. With an increased number of successful samples per token of input, the learner assigns on average less weight to what might be a wrong hypothesis, here the Verb-selecting parameter value. The reason for modeling a more tentative learner is as follows.

I ran three simulations of the model. The simulations differed as to what the probability of modification of a vP shell was, a probability that was held constant throughout the course of each simulation. Recall from the preceding section that it is this probability that decreases the likelihood of the Verb-selecting grammar generating output that matches the input. However, the smaller the probability, the smaller the difference is between the two grammars in generating target output. If the probability becomes very small, then the difference between the two grammars becomes quite slight indeed. In one of the simulations, the likelihood of modification was set to be as low as 1%. Nevertheless, the Root-selecting grammar is still the more probable one, and this becomes evident over the course of a large number of samplings. With an increased number of chews, the learner is more likely to sample and thus reinforce the Root-selecting grammar for each token of input. In other words, so long as the learner is sufficiently tentative, even if the modification is relatively rare, in the long run the learner will get pushed toward the grammar of the subset. Indeed this is what we see in the results given below.

Table 5.1 shows the results of the three simulations, which vary according to the frequency of modification. For each simulation, the model was run 10 times, for a total of 30 runs. The table shows the proportions of a parameter value's pseudo-count total out of 10 (which is the sum of the pseudo-counts of all of a given parameter's values). A high average represents a point in the learner's development at which the model has learned the corresponding parameter value. In all three simulations, the model is highly successful at learning the Root-selecting parameter value. This is true even when the likelihood of modification is very low, at 1%. What we also see is that the average

number of tokens to learn Root-selecting increases as the likelihood of modification goes down. This follows from the discussion above. The smaller the chance of modification, the more likely a Verb-selecting analysis is to generate target output. With a sufficiently large corpus of input, though, on average Root-selecting cause will be chosen to generate target output more frequently, and will emerge as the more grammar of best fit to the input.

Table 5.1 Average proportions of weights for Root/Verb-selecting given different frequencies of modification of vP shells

Likelihood of modification	**Root-selecting proportion**	**Verb-selecting proportion**	**Average tokens of input per 10 runs**
1%	.95	.05	1922
5%	.98	.02	626
50%	.99	.01	82

Further, a look at the developmental course of the model reveals that more generally the model rarely assigns a strong weight to the Verb-selecting value. In general, the weights for the Verb-selecting value go steadily down. On only one run does a weight for the Verb-selecting value rise above 75%. This is during the simulation with a 1% likelihood of modification of a vP shell. On this run the model initially strongly favors Verb-selecting, but then the weight for Verb-selecting cause drops, and the model settles even more strongly on Root-selecting. There is thus almost no categorical variability in parameter setting throughout the course of learning. In sum, the model is consistently learning a Root-selecting grammar, even with a very low likelihood of modification of vP shells.

3.4 Learning the grammar of the superset language

So far I have discussed how the implicit negative evidence of the absence of expected adverbials pushes the likelihood of parametric weight toward the simpler hypothesis and away from the more complex hypothesis. The results in Section 3.3 show for the simple English data that the more complex hypothesis will not be learned, whereas the simpler Root-selecting hypothesis will be. At this point we can now ask whether the more complex hypothesis could ever be learned. In a sense, the learning question that I began the chapter with has been flipped. We began with a question about how to learn the grammar of the subset language. Given that the model is successful in learning the grammar of the subset language, the question then turns to how the Verb-selecting causative can be learned in a language that has such a setting for the Cause-head selection parameter. In this section, I offer some thoughts about what a Verb-selecting language would look like, given the framework adopted here.

Recall that overt evidence that contains the low adverbial reading unambiguously tells the learner that the target grammar is Verb-selecting. However, the approach I took

in Section 3.2 was to attempt to learn the target grammar without input that unambiguously illustrates the high or low readings. This approach was taken out of concern that such input might be sufficiently rare that it would play a minimal (or non-existent) role in the learning outcome. Assuming that such input is vanishingly rare in the learner's input corpus, then the question remains: how does the child learn a target grammar that is Verb-selecting?

Given that the grammar of the subset language is favored in Section 3.2, the model predicts that Verb-selecting languages must have some other (presumably not infrequent) positive evidence that they are Verb-selecting in order to be learned as such. In what follows I propose two ways in which a Verb-selecting language could be learned. The first is based on one of Pylkkänen's (2008) diagnostics for a Verb-selecting causative, and involves identifying phonologically overt verbal morphology between the Cause-head and the root. The second is a possibility that emerges in the model I have proposed and involves input to the learner that contains a high proportion of modifiers of vP shells. Such input is more compatible with a Verb-selecting grammar and could thus push the learner toward adopting the grammar of the superset language.

One kind of evidence is phonologically overt verbal morphology between the causative morpheme and the root. Recall that such intervening morphology is impossible under a Root-selecting hypothesis (there is simply no structural position for it), whereas with a Verb-selecting structure this verbal morphology can be neatly accommodated as different kinds of little-v heads. Pylkkänen discusses Finnish and Bemba as examples of languages that have a Verb-selecting Cause-head. Indeed, both Bemba and Finnish show verbal morphology intervening between the causative morpheme and the verb root (though the frequency of this intervening morphology is not discussed). I illustrate this with two examples from Finnish (Pylkkänen: 116-117):

(22)	a. ravio-	'rage'		b. seiso	'stand'
	ravio-stu	'become enraged'		seiso-skele	'stand around'
	ravio-stu-tta	'cause to become enraged'		seiso-skel-utta	'cause to stand around'

This intervening morphology is thus unambiguous evidence that supports the grammar of the superset language. So long as this kind of evidence is sufficiently robust, the learner can rely on it to learn this grammar without any difficulty.

A second kind of positive evidence that favors the grammar of the supserset language is perhaps more commonplace, although it is ambiguous evidence. As was alluded to in Section 3.2, input with any kind of modifier of a vP shell, though compatible with either parameter value, actually favors a Verb-selecting analysis. The more complex the hypothesized structure is, the more compatible it is with modified input; the less complex the structure, the more compatible it is with non-modified input. This was shown schematically in (18) and (19), repeated below.

(18) *Expectations of modification*

$$\underbrace{p(\textit{Modified-CauseP}) \vee p(\textit{Modified-vP})}_{\text{Verb-selecting}} \quad > \quad \underbrace{p(\textit{Modified-CauseP})}_{\text{Root-selecting}}$$

(19) *Expectations of <u>no</u> modification*

$$\underbrace{p(\neg\textit{Modified-CauseP})}_{\text{Root-selecting}} \quad > \quad \underbrace{p(\neg\textit{Modified-CauseP}) \wedge p(\neg\textit{Modified-vP})}_{\text{Verb-selecting}}$$

In other words, modified input favors the grammar of the superset language, whereas non-modified input favors the grammar of the subset language. With a sufficiently large proportion of the corpus containing modified input, which favors Verb-selecting, the balance of data would shift, and the learner can be pushed toward learning the grammar of the superset language.

The possibility of learning the grammar of the superset language from modified input raises an interesting tension between the two kinds of ambiguous input (i.e. modified and non-modified input). An important step moving forward would be to expand the corpus of the model in Section 3.2 to include these two kinds of input. Future research can look at (a) identifying the range of the proportion of modified input necessary to learn the Verb-selecting grammar, and (b) comparing this range to frequencies found in corpora of child directed speech. How much of the corpus must contain modifiers of vP shells before the learner adopts the grammar of the superset language, and is this proportion ever attested in natural language?

I speculate that at least in the case of English ZDCs, modifiers of vP shells are not superabundantly frequent, and thus that they would not pose a problem in learning the grammar of the subset language. To take an example, let us suppose that the probability of a modifier of a vP shell is .05, and that this probability matches the frequency in the corpus of ZDCs that have a modified vP shell. (Recall from the results in Section 3.3 that the model has no trouble learning the grammar of the subset language with this frequency of modification and only non-modified input.) With two kinds of input, this would mean that the 5% of the time when the learner encounters a modified token of input, the grammar superset language will be slightly favored. In contrast, the overwhelming majority of the time (95%), the grammar of the subset language will be similarly favored.[10] Such a scenario does not present us with an obvious obstacle to learning the

[10] As in (17) and (21), if we assume priors of 0.5 for either Root- or Verb-selecting cause, and assuming a probability of 0.05 for modification of a vP shell, then the probability of generating non-modified output under Root-selecting is .475, and .4512 under Verb-selecting (cf. example (19)). With the same priors, the

grammar of the subset language for English ZDCs. However, this proportion of modified input could be helpful, in conjunction with the unambiguous evidence of intervening verbal morphology of the sort in (22), in learning the Verb-selecting grammar. This would allow for learning the grammar of the superset language with a smaller proportion of the input containing intervening verbal morphology. In other words, the two types of evidence could work in tandem to learn the grammar of the superset language, and neither type of evidence would need to be as frequently attested were it the only kind of evidence to push toward the this grammar.

This section has made a key prediction concerning learning the grammar of the superset language: some other kind of positive evidence is necessary to push the learner away from the grammar of the subset language. I have presented two kinds of positive evidence that could serve this role. Some of the details related to this evidence are admittedly speculative, but they give some perspective on what the learning task involves with respect to learning the less restrictive Verb-selecting hypothesis.

4. Comparison with other models

In this section I briefly compare the probabilistic model proposed here with some other learning models for parameter setting in syntax. I contrast the model here with an alternative probabilistic, Yang (2002), and two non-probabilistic models and Sakas and Fodor (2001) and Gibson and Wexler (1994), all of which struggle with the learning challenge posed here by zero-derived causatives in English.

Recall from Chapter 2 that the core of Yang's (2002) probabilistic learning model involves increasing or decreasing a parameter value's probability based on whether adopting that parameter leads to a grammar that is compatible with the input data. Thus whenever the model encounters any data containing ZDCs, it will sample a Cause-selection parameter value based on the probability distribution and test out this value to see whether it is compatible with the input. The scenario of ZDCs in English, then, is problematic for Yang's model. All the relevant parameter values are compatible with the input, and there is thus no input data that can systematically rule out any of the parameter settings. As both Root and Verb-selecting parameter values will have similar reward-punishment rates in this situation, all things being equal (e.g. non-biased priors), the model could converge on either setting or get stuck in a state of stasis, with neither setting's probability exhibiting asymptotic behavior. These two possibilities are illustrated in the simulations in Pearl (2007), and whether the model settles on a single parameter value is a function of how tentative the learner is. Compared to the generative model proposed here, Yang's model is unable to learn from implicit negative evidence. Yang's model does not go beyond grammar compatibility to consider the probability of the data given a particular grammar. Such a conditional probability allows the model here

---

probability of generating output with a single modifier of a vP shell under Root-selecting is .025, but now .0487 under Verb-selecting (cf. example (18)).

to learn from implicit negative evidence and to evaluate which of multiple input-compatible grammars is a better fit to the corpus of input.

Similarly, in the error-driven model of Gibson and Wexler (1994), there is no guarantee that the learner will converge on the target parameter setting for ZDCs. In this model, parameter settings have weights of 1 or 0, and a parameter's value is changed only if the current vector of parameters is incompatible with the most recent token of input. In such a case, only one parameter can be changed (the Single Value Constraint). Which parameter is chosen to have its value changed is left as an open question, but there is a constraint such that whatever the new parameter vector is, the grammar represented by that new vector must now be compatible with the most recent input (the Greediness Constraint).

Consider, then, how the Gibson and Wexler model fares if the initial state, which is some random grammar or parameter vector, has a non-target parameter setting for English ZDCs. No input containing a ZDC could force the Cause-selection parameter to change its value because both settings are compatible with that data. Further, even if this input forced the model to change its current grammar (because of non-target setting of some other parameter), the model would not change the setting of the Cause-selection parameter because no new value for this parameter would help in the face of the latest input (the Greediness Constraint). The model would have to change the value of some other parameter and leave the Cause-selection parameter alone (the Single Value Constraint). Thus the model will be in a local maximum: no input could push the model toward a target setting for ZDCs, and the model would remain stuck in a non-target setting. Of course, if the initial state was a Root-selecting grammar, then no input in English would push the learner from that setting, and the learner would have the target parameter setting.

Finally, the model in Sakas and Fodor (2001) crucially relies on input that contains unambiguous triggers. In their model, as the parser builds a parse tree of the input, the parser is able to recognize at any point in the structure whether a parametric choice is underdetermined given the input data. For the case of ZDCs discussed in this chapter, the parser, upon facing the Cause-head in the parse tree, presumably would be able to determine that either a vP or √P complement is compatible with the input data. In the terms of Sakas and Fodor, the parser is faced with an ambiguity with respect to parameter setting. What the parser then does is report this ambiguity to the learning mechanism. The learning mechanism will then not use this 'ambiguous input' to learn a parameter setting. In other words, the learning mechanism will wait until an unambiguous trigger occurs in the input before setting any parameter value. Now as we have discussed, all the relevant data for zero-derived causatives in English underdetermine the correct structural analysis – it is all ambiguous input, and there is no unambiguous input. As it stands then, Sakas and Fodor's model is unable to learn the correct parameter setting when faced with the challenge of ZDCs.

Before closing this section, I note that an amendment to both Gibson and Wexler's and Sakas and Fodor's models would be able to account for the Cause-selection parameter: a default parameter setting. The learning mechanism would only need to consider other parameter settings if pushed toward them by the input. If Root-selecting cause is the default value, then the English zero-derived causatives would be accounted for. Only if the input data presented some evidence that is incompatible with a Root-selecting parameter setting (e.g. an utterance with the low adverbial reading) would the learning mechanism change from the default to a Verb-selecting setting. As mentioned in Section 2.2, though, an advantage of the model here is that no default needs to be specified.

Further, assuming a default value for English ZDCs makes the prediction that children never pass through a developmental stage in which they adopt a Verb-selecting grammar (at least some of the time). This prediction follows from the claim that no evidence ever forces the learner to abandon a grammar that initially adopts the simpler causative structure. In contrast, according to the non-deterministic model proposed here, at earlier stages in the learning procedure, non-target parameter settings with likelihoods that are not too low are viable choices. Before parameter setting is finalized, then, we might expect non-target behavior from children with respect to adopting, say, the Verb-selecting grammar. Is there evidence that children sometimes treat zero-derived causatives in English as being Verb-selecting before having learned that they are in fact Root-selecting? The model would lead us to expect that in initial stages of learning, the likelihood of a Verb-selecting analysis is high enough that children would incorrectly treat them as being Verb-selecting at least some of the time. Careful experimental work would be needed to test these predictions, but to the extent that they are borne out, in addition to showing how target parameter settings can be learned, an advantage of the non-deterministic framework here is its potential to model non-target behavior.

5. Summary

In this chapter I presented zero-derived causatives as a case study of a potential Poverty of the Stimulus learning challenge. Can a child reliably learn the grammar of a subset language when it is the target and not the grammar of the superset language? The generative model I have proposed addresses this challenge. Although both grammars are compatible with all the learner's evidence, by being sensitive to implicit negative evidence the model can learn that the grammar of the subset language is a better fit to the input than the grammar of the superset language. Thus there is no poverty of the stimulus, and there is no need for an additional principle such as the Subset Principle: the model is able to converge on the more restrictive grammar through the learning process itself.

Appendix 2: Additional evidence for a Root-selecting grammar in English?

Pylkkänen (2008) claims that the absence of 'causativized unergatives' in English is additional evidence in favor of a Root-selecting parameter setting. Thus if *cry* is unergative in (23a), the ungrammaticality of (23b) putatively shows that unergatives cannot undergo causativization. The reason Pylkkänen gives for this seems to be the following: if the *crier* is introduced by Voice in (23a) it must also be introduced by Voice in (23b), an impossibility given that the Cause-head is Root-selecting and allows for no verbal morphology such as the Voice-head between itself and the root.

(23) a. John cried.
b. *Mary cried John. ('Mary caused John to cry.')

Now if Pylkkänen is correct, this would be useful for the model in Section 3. The learner would expect lots of utterances of the form in (23b), but as there are none in the primary linguistic data, this would constitute another source of implicit negative evidence to learn from. The model would then learn from two sources of implicit negative data with the potential for even more efficient or expeditious learning.

However, some roots that do appear in unergative environments can also be causativized despite the widespread claim that this is not possible in English. I give two examples here in (24) and (25).

(24) a. John choked
b. {Mary, A fishbone} choked John.

(25) a. The cattle grazed during the day.
b. During the day the men drove and grazed the cattle and at night herded them by relays.
(Accessed 23 Nov. 2013: http://www.co.wilbarger.tx.us/CattleDrives.htm)

Note that both verbs pass the *x's way* test that Levin and Rappaport-Hovav (1995) give as a diagnostic of unergatives in (26), and fail to license resultatives in (27), which Levin and Rappaport-Hovav discuss as a diagnostic for unaccusatives.

(26) a. …as he choked his way *(out of the waves). (directed motion reading)
(Accessed 23 Nov. 2013: http://www.goodreads.com/quotes/520050-there-was-a man-here-lashed-himself-to-a-spar)

b. By mid-afternoon, the band [of sheep], which naturally eats weeds most animals avoid, has grazed its way *(across the hill). (directed motion reading)
(Accessed 23 Nov. 2013: http://www.montana.edu/news/1206/across-montana-sheep-make-short-work-of-weeds)

(27) a. *John choked breathless. (‘John was breathless as a result of choking.’)
b. *The cattle grazed exhausted during the day.
(‘The cattle were exhausted as a result of grazing.’)

I conclude that these are counter-examples to Pylkkänen’s claim.

Further, although there are many impossible ‘causativized unergatives’, one of which is illustrated by (23). There are also many impossible ‘causativized unaccusatives’, such as (28) as discussed by Blanco (2010).

(28) a. John arrived.
b. *Mary arrived John. (‘Mary caused John to arrive.’)

An important research question that I do not address here is the appropriate characterization of which roots can appear in both causative and unergative/unaccusative environments (though see Ramchand 2008 for a proposal). Nevertheless, I conclude on the basis of (24)-(27) that until more is known about the differences between examples like (23) and (24)-(25), examples such as (23) cannot be used as evidence in favor of a Root-selecting parameter setting.

Chapter 6

Further Discussion: Learning biases

1. Introduction

In previous chapters I have taken the general approach that learning happens without biases. We have seen that my starting assumption has been that the priors for parameter values are equal and thus non-biased. Further, nothing constrains which grammar is learned beyond the shape of the input. Thus although parameters interact in finding the grammar of best fit to the input, the learner is not predisposed to learn any particular kind of grammar: learning one parameter value does not intrinsically bias the learner toward learning some other parameter value. This general approach can be considered the null hypothesis. Further, this approach has had its successes as has been illustrated in the proof-of-concept simulations in Chapters 3-5. More generally, though, we can ask whether this is the right approach for language acquisition. To address this question we can ask what the motivation might be to depart from the null hypothesis. In this chapter I discuss two possible alternatives to the null hypothesis by picking up the thread of two discussion points that have been raised in earlier chapters: the inadequacy of a universal set of defaults in a deterministic learner and how to model language universals. As we shall see, there is evidence against a deterministic learner with defaults, which constitute biases for certain parameter values. However, a learning bias concerning the relations between different parameters may help the model I have proposed capture certain language universals. Before investigating these points further, let us see how these alternatives relate to learning biases and why one might want to pursue them as alternatives to the null hypothesis.

First, let us consider the issue of a universal set of default parameter values.[1] A default value constitutes a bias for that particular value. Unless presented with sufficient evidence to the contrary, the learner will adopt whichever particular grammar is represented by the default. As discussed in Chapters 3 and 4, a default can be implemented probabilistically with one parameter value having a very strong weight in the initial state as a prior. This is true regardless of whether the model is deterministic (and learns only from unambiguous evidence) or not. If the weight is sufficiently strong, then it represents a learner that already has a parameter setting for that value in the initial state. And if the target grammar instantiates the non-default value, then given sufficient evidence in favor of this value, the learner will be able to redistribute the weight so as to reset the parameter. In contrast, with equal priors the learner has equal expectations

[1] Recall from Chapter 1 that I assume the least restrictive view of what a universal default value might be: any of a parameter's values might be the value that all learners use as a default.

regarding any of a parameter's values being the target value and has an equal likelihood of adopting any of them in the initial state.

But why might the learner be equipped with defaults? In probabilistic terms, why pursue an approach in which the priors are biased? In Chapter 1 I discussed how the learning task can be facilitated by a model that learns only from ambiguous evidence (cf. Sakas and Fodor 2001). Such a model does not run the risk of being unduly influenced by ambiguous evidence, and can avoid some disastrous learning error from which in cannot recover. However, I also discussed several immediate difficulties for such an approach to learning, including how to learn when there is insufficient unambiguous evidence and how to model learning errors that indicate a non-target parameter value. In what appears to be the most promising way suggested in the literature to shore such an approach, and therefore attempt to address these issues, a deterministic learner can be equipped with a set of default parameter values (cf. Fodor 1998; Sakas and Fodor 2001; and Sakas and Fodor 2012). Even with defaults, though, such a learner cannot account for the full range of parameter-setting errors we see in learners. In Section 2 I review in more detail an argument first introduced in Chapter 1 against a universal set of defaults. The argument is based on the inadequacy of a universal set of defaults in a deterministic learner to account for the error patterns we see in children cross-linguistically. A deterministic learner with defaults predicts only one kind of error: an error that reflects the default value before the child has learned the target setting. I will review some evidence from Swedish and English that children learning these languages have errors that reflect both values of a binary parameter. We can make sense of these error patterns if the children are learning from ambiguous evidence with equal priors, and I sketch what the learning scenario looks like that would result in these errors. In short, there is a substantial amount of ambiguous evidence that is compatible with the non-target parameter value in the two languages (indeed sometimes this evidence actually favors the non-target value). Further, with equal priors, the learner has the greatest likelihood of choosing either parameter value when given ambiguous evidence. This maximizes the likelihood the model can successfully account for contrasting errors patterns where in one language on the basis of ambiguous input the learner can be influenced to adopt one non-target value, while in another language ambiguous input might lead to the other non-target value. In the absence of any evidence to the contrary, then, I will assume in general that the null hypothesis of equal priors holds in the learning model I have proposed (though see the discussion of object raising in Chapter 4 for an exception).

I now turn to the issue of language universals, in particular how the learner can account for grammars that are not attested cross-linguistically. In Chapter 4 I discussed how a gap in the inventory of the grammars of the world's languages might not be accidental. Rather, this gap might arise because the unattested grammar is very hard to learn (possibly unlearnable). In particular, I focused on the Final-over-Final Constraint (FOFC) from Biberauer et al. (2014), which describes a putative language universal:

certain combinations of parameter values for different parameters of head-complement order are not attested. For example, the claim is that in no language does a head-final TP take a head-initial VP as its complement. We saw that the learning model cannot currently capture the idea that a FOFC-violating grammar is hard to learn. The trouble for the model is that nothing in principle constrains the model from reinforcing parameter values, for example [T-fin] and [V-init], above and beyond fitness to the input. Thus if [T-fin] is a good fit to the input, and if [V-init] is a good fit to the input, nothing prevents the model from learning a [T-fin, V-init] grammar. But what if there were a bias against learning such a grammar? What if some bias helped regulate the relations between different values that are sampled and reinforced? In Section 3 I sketch out a proposal for such a bias. According to this proposal, there is a bias against reinforcing certain values in a FOFC-violating grammar, whereas no such bias exists in a FOFC-compatible grammar. This bias is a departure from the null hypothesis, according to which all parameter values are reinforced equally in a particular sampling of values that fits the input. But if successful, such a bias would be a reason to depart from the null hypothesis, as it would extend the empirical coverage of the model by allowing it to capture a putative language universal such as FOFC. It would also lay the groundwork for capturing any other universals or cross-linguistic tendencies that involve relations between multiple parameter values.

2. A problem for defaults

I have just discussed how a deterministic learning model provides some conceptual motivation for a learning bias in the form of a universal set of default parameter values. However, in Chapter 1 I introduced the idea of contrasting error patterns as being a challenge for a deterministic learner that relies on such defaults. Contrasting errors patterns describe the following scenario involving child errors. Suppose there are two languages with different target values for a particular binary parameter. Further suppose that children learning those languages make errors that illustrate a non-target value for that parameter. The non-target value will be different for each language, giving us a situation of contrasting errors. For a deterministic learner with defaults, an error of a non-target parameter value must be the result of a default value. But no matter what the default is, such a default value cannot account for both kinds of errors we see with contrasting errors. In this section, I review in more detail an example of contrasting errors in English and Swedish children. I then sketch how a model that learns from ambiguous evidence with equal priors could account for these error patterns. In the context of this chapter, this section thus provides an argument against a particular implementation of a learning bias, namely the use of defaults as part of a deterministic learner. To the extent that this argument holds, then our reason to pursue an approach with non-biased priors as a general property of parameters disappears. Beyond the context of this chapter, this section fleshes out an argument against deterministic learning models introduced in

Chapter 1 (and more specifically, an argument against an attempt to shore them up with default values in the face of learner errors), and as such is a more detailed continuation of the discussion in Chapter 1.

I begin with a short review of defaults in a deterministic learner. Recall that a deterministic learner uses only unambiguous evidence to set parameters. With Korean in Chapter 4 and English zero-derived causatives in Chapter 5, we saw examples of parameters for which there is no or insufficient unambiguous evidence available to the learner. To ensure that all learners can nevertheless arrive at the target grammar, I have suggested that a deterministic learner be equipped with a universal set of default values for all parameters. Indeed, this is the approach that Sakas and Fodor (2012) have adopted for their deterministic learner because of similar considerations. Thus if the learner never encountered unambiguous evidence for a particular parameter, the default value would continue to be used for that learner's grammar.

I have also pointed out that even when augmented with a universal set of default values, a deterministic learner cannot account for variable parameter setting across speakers. Again in Chapter 4 we saw that some learners of Korean adopt a grammar with verb movement, whereas others do not. This variability was possible, in part, because there was insufficient unambiguous evidence in favor of movement or keeping the verb in-situ. Given this lack of unambiguous evidence, if learners had a default to not move the verb, then we would not expect any of them to learn a parameter setting for verb movement. And conversely, if the default value were set positively for verb movement, we would not expect some learners to adopt a grammar that had no verb movement. Let us now set aside the problem of variability for a deterministic learner in order to consider another problem, that of learners' contrasting error patterns.

At first, it appears that a universal set of default values helps a deterministic learner account for learner errors. If a language has unambiguous evidence for only one value of a binary parameter, then learner errors that result from the non-target value must be the result of a default value. This is because in such a language, a model that learns only from unambiguous evidence can only move away from a non-target state (e.g. a non-target default value) and toward a target state. We saw in the discussion in Chapter 3 that this approach is in principle able to capture the errors of the Swiss German children. There I claimed that children initially have a non-target value of T-initial, which resulted in non-verb-final embedded clauses. This is also possible under a deterministic learner that has a default value of T-initial. This value would then gradually get reset to T-final by means of unambiguous evidence as the children learn the adult grammar. The prediction, then, is that children learning some other language will never incorrectly adopt a T-final grammar if the adult grammar of that language is T-initial. Coupled with the errors in Swiss German, T-final errors would constitute a contrasting error pattern, which should not be possible given the discussion above about defaults in a deterministic learner. I am not aware of contrasting errors involving T-initial and T-final for the TP-

headedness parameter, however in what follows I will show an example of contrasting errors involving verb/auxiliary movement.

2.1 Errors in Swedish

I now present some evidence for contrasting error patterns concerning T-to-C movement in the spontaneous productions of children learning Swedish and English. I begin with Swedish.

Like Swiss German, Swedish is a V2 language in matrix clauses, and a variety of different constituents can appear in initial position. Unlike Swiss German, which has an OV order in the VP, Swedish is VO. Following Platzack (1986: 198, 210), we can see that in Swedish the finite verb moves to C if we assume that negation occurs in a fixed position. Thus we see that in (1a) the finite verb precedes negation, indicating that it has raised out of the VP. Depending on the position of negation, different possible landing sites of this movement are T or C, as schematized in (1b, c).

(1) a. Författaren skrev inte någen bok i år.
author.the wrote not any book in year
'The author didn't write any book this year.' (Waldmann 2011: 332)

b. [$_{CP}$ C [$_{TP}$ V_{fin}+T [Neg … ~~V_{fin}~~ …]]]

c. [$_{CP}$ V_{fin}+T+C [$_{TP}$ Neg [$_{TP}$ ~~V_{fin}+T~~ [… ~~V_{fin}~~ …]]]]

Platzack notes, though, that in an embedded context, the finite verb follows negation, as in (2a).[2] We can account for (1a) and (2a) if (a) Swedish has T-to-C movement, which is

[2] Here, I set aside cases of embedded V2 in Swedish, which can occur under certain verbs with an overt complementizer (Holmberg *in press*). Embedded V2 is certainly compatible with the current account of verb movement in Swedish, if the correct analysis of embedded V2 involves stacked CPs (cf. Platzack 1986). As Swedish is [+TC], then if a given complementizer embeds a CP, the finite verb will raise to this lower C-head, SpecCP of the lower C-head could be filled, and the result is a V2 embedded clause. Still, embedded V2 presents several additional learning challenges. First, the learner must identify when an overt complementizer allows for embedded V2, and when it does not. This amounts to learning whether the complementizer embeds a CP or TP – that is, whether the complement is a simpler or more complex structure. I note that this kind of scenario appears to be another candidate for learning from implicit negative evidence along the lines in Chapter 5, in which I considered a similar learning scenario involving competing grammars with structures of varying degrees of complexity. Second, the possibility of embedded V2 could make it more challenging to ascertain the target position of negation, as negation can precede or follow the embedded finite verb depending on whether or not there is embedded V2 (cf. Waldmann 2014 for discussion of potential errors concerning the position of negation in embedded clauses in Swedish children, and see note 3 for more on learning the position of negation). If the learner can determine whether an embedded clause has stacked CPs, though, a single consistent analysis of the position of negation emerges, as per the discussion above. I leave the proper treatment of these challenges as a topic for future research. For the purposes of discussion, I will assume that a deterministic learner such as in Sakas and Fodor (2001) and the learning model I have presented can both learn the target parameter settings in Swedish, including those involved in embedded V2. My focus in the text, rather, is to show that certain errors are not expected under a deterministic learner, whereas they can plausibly be understood as resulting

blocked by the overt complementizer in (2a); and (b) negation in Swedish is higher than T. Under this analysis, the finite verb moves to C in matrix clauses, as in (1c), and precedes negation. According to the kinds of parameter spaces I have been adopting in this thesis, verb movement to C is implemented by positively set parameters for obligatory V-to-T movement [+VT] and for obligatory T-to-C movement [+TC]. V-to-T movement is seen in embedded clauses with an overt complementizer, which blocks subsequent movement to C, as in (2b). Because negation is higher than T, the raised verb will follow negation in these embedded contexts.

(2) a. Vi frågade [om författaren inte skrev någen bok i år].
we asked if author.the not wrote any book in year
'We asked if the author didn't write any book this year.' (Waldmann 2011: 332)

b. …[$_{CP}$ if [$_{TP}$ author [$_{TP}$ Neg [$_{TP}$ wrote+T [… ~~wrote~~ …]]]]]

From a learning perspective, the learner can arrive at a [+TC] grammar after encountering evidence of the sort in (1a) and (2a). If we use the same hypothesis space as in Swiss German, with the same parameter settings for matrix and embedded clauses, then the only grammar that is consistent with all this evidence is one in which the verb moves to C (again, assuming a single position for negation). Thus along the lines of the discussion of Swiss German in Chapter 3, with the probabilistic learner I have proposed, a [+VT, +TC] grammar can emerge as the grammar of best fit because it is compatible with all the evidence we have discussed. Moreover, for a deterministic learner, there is unambiguous evidence for a [+VT, +TC] grammar. If a deterministic learner has correctly set [+VT] via unambiguous evidence, such as (1a), then the learner can unambiguously fix the position of negation as being above T with (2a). Consequently, matrix input such as (1a) becomes unambiguous evidence for [+TC].

Turning now to the acquisition data, Waldmann (2011) presents corpus evidence from a Swedish child Tea (ages 1;6-4;0) that shows that the finite verb is sometimes in T and not in C. That is, Swedish provides us with learner errors for a non-target value of T-to-C movement, here a non-target value of [–TC]. The sort of evidence that clearly shows this is V3 productions of the sort given schematically in (3a). What an utterance like (3a) shows us is that the finite verb has moved out of the VP because the verb precedes some constituent ZP, which c-commands the VP. If ZP, say, were not present in (3a), then the utterance is ambiguous: the verb may have raised out of the VP, or it may have remained in-situ. Further, if we assume that SpecCP is a privileged position, and that only one constituent can adjoin to CP, then the fact that two constituents precede the raised verb indicates that the finite verb has not moved to C. I follow Waldmann in assuming that the

---

from learning from ambiguous evidence under the learning model I have proposed. As far as I can tell, this point is independent of the proper treatment of embedded V2 in Swedish.

verb in (3a) has thus moved to T. The schematic structure of an utterance like (3a) is given in (3b).

(3) a. # XP YP V_{fin} ZP [$_{VP}$ …]
b. [$_{CP}$ XP [$_{CP}$ C [$_{TP}$ YP [$_{TP}$ V_{fin}+T [ZP … ~~V_{fin}~~ …]]]]]

Thus utterances of the form in (3) are taken to provide unambiguous evidence that the learner is using a [+VT, –TC] grammar: the finite verb moves to T, but not to C.

Waldmann (2011) reports a number of V3 productions like (3) in Tea's corpus up to around age 3;6. Of 603 utterances that are not subject-initial and that have a finite verb, 185 or 30.6% have the finite verb in third position. This is a fair amount of evidence that the verb is not in C, although these utterances are ambiguous as to the structural position of the verb. The verb could have raised as high as T, which would mean the learner has a [+VT, –TC] grammar, or the finite verb could have remained within the VP. According to the discussion in Chapters 3 and 4, this latter possibility is actually compatible with two relevant types of grammars: (a) [–VT, –TC] or (b) [–VT, +TC]. Under both grammars, the verb remains in-situ, but it is unclear whether T has moved to C. For the purposes of comparison with English, we are interested in learner errors involving a non-target [–TC] grammar. It is likely that a number of V3 utterances are the result of a non-target [–TC] grammar, but it is possible to be more precise with unambiguous evidence like (3) for such a grammar. Productions like (3) will be a subset of all V3 utterances because they involve an additional constituent that specifically allows us to determine the lower boundary of the verb's position. To identify this lower boundary, Waldmann relies on productions that contain negation that follows the verb. If the verb precedes negation, then it can be taken to have moved out of the VP, as per the discussion of (1). Thus negation is an example of ZP in (3). Waldmann (2011: 13) reports that Tea has 13 productions like (3) with negation, which unambiguously illustrate a [+VT, –TC] grammar. Some examples are given in (4).[3] The bulk of these productions occur between ages 2;10-3;3.

[3] Interestingly, given the discussion of the adult grammar of (1) and (2), we are now left with the conclusion that Tea also has non-target placement of negation below T in the examples in (4). Recall that if we follow Platzack (1986), then negation occurs above T in the adult grammar. Here we are using the examples in (4) to show that Tea has raised the verb to T, yet the verb precedes negation. If negation were correctly merged higher than T, we would expect the verb to follow negation in (4). If this discussion is on the right track, then the non-target placement of negation in Tea's productions is an additional error that warrants further investigation. The error is perhaps not surprising in light of Waldmann (2014), who presents a detailed discussion of Swedish children's errors concerning the position of the finite verb with respect to negation in embedded clauses (see below for more discussion of Waldmann). I will not engage in further investigation here, though. For our purposes, it is sufficient to observe that Tea has a non-target [–TC] grammar. Moreover, there is no obvious causal relationship between misplacing negation and not raising the verb to C. As such, the non-target placement of negation is orthogonal to the discussion at hand. Nevertheless, it is still possible that while one of these errors does not cause the other, these errors might still be traced back to a similar source. What we can point to is the observation that, as discussed in the text, (1) and (2) on their own constitute ambiguous evidence as to the position of negation (cf. also note 2 for

(4) a. # domma Bella har inte. (Tea: 2;03)
them Bella has not
'Bella doesn't have them.'
(cf. *Dess har Bella inte*.) (Waldmann 2011: 347)

b. # Nu dom öve inte leta efter det. (Tea: 2;11)
now they need not look for it
'Now they don't need to look for it.'
(cf. *Nu behöver dom inte leta efter det*.) (Waldmann 2011: 341)

Thus we have seen that Swedish has child errors indicating a non-target value of [–TC]. This contrasts with what we will see in English. English is a [–TC] language, but there is evidence for child errors indicating the use of a grammar with a [+TC] parameter value. Together, the Swedish and English errors constitute the kind of contrasting error pattern that is problematic for a deterministic learner with defaults.

2.2 Errors in English

Before discussing the child errors in English, I first review the target parameter values in the adult grammar (cf. Pollock 1989 for a similar outline of verb movement in English). In English, verb movement is restricted to finite auxiliaries (including the copula and modals), whereas in Swedish verb movement can apply to any finite verb form, whether an auxiliary or a main verb. Consequently, any discussion of the verb movement parameters for English must be understood as being relevant only for auxiliaries. Nevertheless, an error with respect to the placement of an auxiliary in T or C in child English is still an error concerning T-to-C movement. I will assume that the same basic parameter for T-to-C movement needs to be set in Swedish and English. A child learning English must certainly learn that auxiliaries pattern differently from main verbs

---

more on potential ambiguity concerning the position of negation). That is, both data points in isolation are compatible with negation being above or below T. Similar ambiguity exists for the position of the verb in these utterances: is it in T or C? In the text below, I will suggest that this ambiguity plays a role in the [–TC] errors. I note that this could also play a role in the non-target placement of negation.

I refer the reader to Waldmann 2014 for an alternative approach to the non-target placement of the finite verb and negation in the Swedish children's embedded clauses. A detailed discussion of Waldmann would take us too far afield, but Waldmann's proposal bears some similarity to the approach I have been advocating, according to which the errors can be attributed to the influence of ambiguous evidence (cf. Waldmann 2014: 62). One core difference regards the position of the verb in embedded clauses, which Waldmann assumes remains in the VP (Waldmann 2014: 64). Presumably for Waldmann the distribution of the verb in the target grammar would result from a [+V-to-C, –V-to-T, V-init] grammar. Such a grammar gives us V2 in matrix clauses, but embedded clauses with a complementizer block movement to C. Further, because of the [–V-to-T] setting, the embedded verbs remain in-situ and thus follow negation (but precede the object). As discussed in Chapter 3 (note 21), though, it remains an open question as to whether a hypothesis space with a parameter for V-to-C movement can adequately account for the Swiss German children's errors. If it does not, then it presents a challenge to Waldmann's analysis, which would seem to rely on the inclusion of such a parameter. As this question cannot be resolved here, I will simply note that Waldmann's approach as I have described it here remains plausible, but that I will not pursue it any further here.

in English, but this point is orthogonal to the general point here of illustrating contrasting errors, and I will abstract away from it.

Let us now consider the movement properties of auxiliaries in the adult grammar of English. First, I assume that the finite auxiliary moves to T. Evidence for this comes from the fact that only the finite auxiliary can precede sentential negation, as shown in (5a-c). If negation marks the edge of some lower verbal domain, then we can account for the distribution in (5a-c) if the finite auxiliary moves higher than negation to T. This is presented schematically in (5d).

(5) a. Mary (*not) has (not) swum today.
b. Mary (*not) is (not) swimming now.
c. Mary (*not) has (not) been (*not) swimming for long.
d. [Aux_{fin}+T [not [~~Aux_{fin}~~ […]]]]

If the finite auxiliary moves to T in English, then we can use V3 utterances to conclude that the auxiliary does not generally move to C. Some examples are given in (6), which are English versions of (3). (6a) illustrates this in a matrix clause, and (6b) shows V3 in an embedded *wh*-clause. If, for example, the subject in (6a) were in SpecCP, the presence of the adverb *probably* following the subject shows us that the auxiliary has not raised to C.

(6) a. Mary probably has (not) swum today.
b. I know [which homework assignment Mary has not done yet]!

We can conclude, then, that English in general is a [–TC] language. Nevertheless, in some kinds of clauses we do see movement of the auxiliary to C. I will focus just on the case of interrogatives here. In interrogative clauses, such as (7a), we now see subject-aux inversion. This is evidence for movement of the auxiliary to C, which is schematized in (7b).

(7) a. Why has Mary (not) done her homework assignment?
b. [$_{CP}$ why [$_{CP}$ Aux_{fin}+T+C [$_{TP}$ Mary ~~Aux_{fin}+T~~ …]]]

Examples such as (7a), which have T-to-C movement, contrast with embedded *wh*-clauses that lack such movement: the former are interrogative clauses, whereas the latter are not; crucially it is interrogatives that lead to auxiliary movement.

To capture the properties of auxiliary movement to C in English, I will adopt the parameters that Sakas and Fodor (2012) propose for such cases. Sakas and Fodor propose that in addition to a general T-to-C movement parameter that applies across-the-board in all clauses, there can be a more specific parameter concerning T-to-C movement just in questions. This is given in (8), which can be compared with the more general parameter

in (9). The value [–TC] is to be interpreted such that T does not move to C unless some other parameter value, such as [+TC-Q], allows it to do so.

(8) *T-to-C-Q movement parameter (T-to-C movement in questions)*
a. [+TC-Q]: T moves to C in questions if no free morpheme in C blocks it
b. [–TC-Q]: T does not move to C in questions

(9) *T-to-C movement parameter (generalized T-to-C movement)*
a. [+TC]: T moves to C in all clause types if no free morpheme in C blocks it
b. [–TC]: T does not move to C in all clause types

The parameter in (8) is a simple way of capturing the intuition that the properties of C are different in various types of clauses. Moreover, the parameter in (8) provides a straightforward way of implementing the acquisition of T-to-C movement in a learning model. In English, the target grammar is [–TC, +TC-Q]. Unambiguous input such as (6a) can be used by the learner to support the hypothesis that English is [–TC]. If the learner has set [–TC], then input such as (7a) provides unambiguous evidence that English is [+TC-Q].

Including the parameter in (8) for [±TC-Q] does not affect the overall picture for the end-state of a child learning Swedish. In Swedish, which has V2 in both declaratives and interrogatives, the value for [±TC-Q] is actually irrelevant in the learner's final state. Even if a Swedish child were to acquire a [–TC-Q] grammar (by, say, never changing a default value), it is sufficient to have correctly set [+TC] for the more general parameter (9) to arrive at the target grammar with T-to-C movement in both declaratives and interrogatives. Setting the value for [+TC] can done using evidence such as (1) and (2) as has been discussed above. Further, the parameter for [±TC-Q] does not tell us anything about the errors in child Swedish we saw above in (4). Those errors occurred in declaratives. Accordingly, the value for [±TC-Q] is not relevant in the analysis of these errors, and these errors are still attributable to a non-target value of [–TC]. Thus a parameter for [±TC-Q] is fully compatible with the discussion above concerning T-to-C movement (or lack thereof) in both adult and child Swedish.

I have now addressed parameterization of English and Swedish to account for the basic facts of T-to-C movement in both languages. This involved introducing a new parameter to account for the more nuanced distribution of auxiliaries in English. This parameter concerns T-to-C movement in a subset of clauses, viz. interrogatives, and the effect of this parameter will be masked at the end-state of learning if a language has T-to-C movement more generally, as in Swedish.

We have also seen that child Swedish provides evidence for errors of non-target [–TC] in a [+TC] language. I turn now to the case of errors in child English, which provide evidence for [+TC] errors in a [–TC] language.

Stromswold (1990) is a detailed examination of auxiliaries in the corpora of spontaneous productions of 14 children learning English. What is of interest here is

whether the children have subject-aux inversion errors in embedded *wh*-clauses. In the target grammar, inversion is generally not possible, and this is attributed to these clauses not being questions. However, children produce embedded *wh*-clauses that contain non-target inversion. Stromswold (1990: 161) scored 364 embedded *wh*-clauses and found that 36 of them (9.8%) had a non-contracted auxiliary or copula that preceded the subject. Some examples are given below.

(10) a. # No let me see [who is that]. (Nina 2;10)
b. # They don't say [where is my Great Pumpkin book] any more. (Ross 3;08)
(Stromswold: 296)

What is the analysis of these subject-aux inversion errors? Let us assume that the children's productions of these embedded *wh*-clauses are indeed not interrogatives. Accordingly, the value of the TC-Q parameter is not at issue: a non-target value for the TC-Q parameter could result in a production error only in interrogatives. I take the errors in (10) to be evidence for a non-target [+TC] value that is occasionally adopted by English children. Under the target [+VT, –TC] grammar, the auxiliary/copula would remain in T and follow the embedded subjects, which are in SpecTP. With a non-target value of [+TC], the auxiliary will raise past the subject, resulting in the errors we see in (10). The proposed structure of the examples in (10) would then be as in (11).

(11) #… [$_{CP}$ *wh* [$_{CP}$ Aux$_{fin}$+T+C [$_{TP}$ Subj [$_{TP}$ ~~Aux$_{fin}$+T~~ [~~Subj~~ ~~Aux$_{fin}$~~ …]]]]]

Is there an alternative analysis of the inversion errors in (10)? I briefly consider two related possibilities but conclude that they are unlikely to rule out the analysis of [+TC] in (11) for all cases of embedded inversion. One possibility, sketched in (12a), would be that the auxiliary does not raise to T, and the subject remains in-situ in the lexical predicate of the embedded clause (i.e. the subject does not raise to SpecTP). A variant of this approach is to suppose that the auxiliary does raise to T, but that the subject still remains in-situ (12b).

(12) a. # *Subject in-situ; Aux does not raise to T*
#… [$_{CP}$ *wh* [$_{CP}$ C [$_{TP}$ T [Subj Aux$_{fin}$ …]]]]

b. # *Subject in-situ; Aux raises to T*
… [$_{CP}$ *wh* [$_{CP}$ C [$_{TP}$ Aux$_{fin}$+T [Subj ~~Aux$_{fin}$~~ …]]]]

The structures in (12) do not involve T-to-C movement while still giving us the non-target word order found in (10).

However, corpus evidence suggests that the structures involved in these alternatives to the errors in (11) are rarely produced. Pierce (1992) provides corpus evidence from 4 English children that shows that they are highly successful at raising the subject. At first,

it would appear to be difficult to determine whether children raise subject. For example in a simple SV(O) utterance, it is not clear whether the subject remains in-situ or has raised to SpecTP. The evidence from Pierce comes in a somewhat indirect form and is based on looking at non-target child productions that have the order #VS. Pierce (1992: 27) scored 60 examples of the subject occurring after a lexical verb. 45 of these (75%) occurred with unaccusative verbs. Some examples are given in (13). The remaining 15 errors involved unergative verbs, 10 of which were produced by a single child; the other children produced only 1 or 2 errors with an unergative verb. Moreover, no post-verbal subject are attested with a transitive verbs – that is, #VSO is not attested.

(13) *Some %VS productions: Peter (1;11-2;3)*
a. # go two bolts
b. # broken the light (2)
c. # comes me!

(Pierce 1992: 24)

Pierce concludes that the children are fairly successful at distinguishing unaccusative predicates from unergative ones. The fact that the children are much more likely to have a post-verbal subject with an unaccusative is taken as evidence for the children treating the subjects of unaccusatives as internal arguments and merging them into the structure as the complement of V (cf. Perlmutter 1978). The errors in (13) would then illustrate a failure to raise the subject by leaving it in its base position.

Crucially, the errors in (13) are very rare. Pierce (1992: 25, 29) found that VS errors occurred in less than 1% of utterances with a verb from a sample of thousands of sentences. Further, on average 23% of each child's verbs were classified as being unaccusative (Pierce 1992: 27). Let us follow Pierce in assuming that children merge the subject of an unaccusative verb as the complement of V. Let us also follow Pierce in assuming that the errors in (13) are the result of not raising the subject. This leads to the following conclusion. The children are systematically raising the subject of unaccusatives hundreds of times at a near-ceiling rate. If we accept Pierce's conclusions, then this is strong evidence that the children are highly successful at subject raising with unaccusatives. By hypothesis, the children are raising the subject to SpecTP, as in the adult grammar. It is less clear whether the children are raising the subjects of unergative or transitive verbs, as the base position of these will precede verb. As the null, hypothesis, though, I will assume that the children are fairly systematically raising the subjects of all predicates to SpecTP.[4]

[4] Puzzlingly, Pierce (1992) concludes that children often do *not* raise the subject to SpecTP. Pierce's claim is largely based on two observations, one involving negation and the other clause-initial auxiliaries. As regards negation, Pierce observes that for young children, productions in which negation precedes a non-contracted auxiliary are virtually unattested (only one example was found; Pierce 1992: 59). That is, we almost never see utterances of the form in (i):

The above discussion has attempted to show that several English children rarely fail to raise the subject. If English children more generally rarely leave the subject in-situ, then it is unlikely that the alternative analyses in (12) fully account for all of the subject-aux inversion errors that Stromswold (1990) identified in embedded *wh*-clauses, as in

---

(i) *Unattested*
#… no(t) Aux Subj / no(t) Subj Aux
#… no(t) Kitty is sleeping / no(t) is Kitty sleeping (Déprez and Pierce 1993: 37)

Pierce takes the absence of (i) to indicate that the children have adult-like placement of negation below T. However, Déprez and Pierce (1993: 37) provide only one example from around this period of development in which an auxiliary precedes non-contracted negation, and Pierce (1992: 59) suggests that this is rare. This observation weakens Pierce's claim about children having target-like placement of negation (cf. below). (Interestingly, Déprez and Pierce's single example does appear to show that the subject has not raised, a possibility I have not excluded, especially given (13), merely one that I have claimed is not common.) If we accept Pierce's claim for the moment, it can be taken to indicate that the children are adult-like in raising the auxiliary to T at a relatively high proportion of the time. Additionally, very early in development, negation frequently appears clause-initially before the subject, resulting in non-target productions such as (ii).

(ii) # No Fraser drink all tea. (Eve: 1:09)
(Pierce 1992: 57)

Pierce then tries to relate (i) and (ii) in the following way. Pierce assumes that the there is a single position for negation, namely the position it appears in in the adult grammar below T, as suggested by (i). If the negation in (ii) is also in the same position below T, then the subject in (ii) must not have raised. However, Pierce's conclusion about subjects generally not raising with clause-initial negation is not straightforwardly compatible with the robustness of subject raising discussed in the main text. Conversely, the evidence I presented for subject raising is not compatible with Pierce's claim that the children have adult-like placement of negation. Indeed, several of the children discussed in Déprez and Pierce (1993), who looked at 3 of the child corpora investigated in Pierce (1992), appear to nearly systematically use clause-initial negation early on in development (up to around age 2). For Pierce, this systematic use of clause-initial negation means that a co-occurring overt subject is in-situ.

It is not obvious, though, that children always have target-like placement of negation in English (cf. notes 2 and 3 for a related suggestion about negation in Swedish). It is plausible that there is more than one position for negation during the course of the children's development. That is, it is possible that the negation in (ii) appears in a position above T (cf. Klima and Bellugi 1966), and that the subject is moving to SpecTP. Although I think this alternative involving multiple positions of negation is plausible, I have no fully worked out explanation for why children would have such an analysis of negation. One speculation is that as other quantificational elements (e.g. 'only', 'every', the 'no' of the nominal domain, etc.) occur overtly in multiple positions (including sentence-initial position) and can plausibly take scope in multiple positions, children might overgeneralize the range of options available for sentential negation. Learning from implicit negative evidence about where negation does not occur (along the lines of Chapter 5) could allow children to learn the correct position for sentential negation.

Pierce (1992) also observes productions of non-target subject-aux inversion in matrix clause declaratives, as in (iii), and also attributes this to the subject remaining in-situ.

(iii) # Is kitty sleep. (Naomi)
(Intended interpretation: declarative) (Pierce 1992: 79)

Again, Pierce's claim does not exactly square with the conclusion of widespread subject raising in the text. In contrast, examples such as (iii) can be taken as further evidence for the idea I have been pursuing in the text, namely that English children sometimes adopt a non-target [+TC] grammar. Accordingly, in (iii) the subject has raised to SpecTP, and the auxiliary has moved to C.

(10). If this reasoning is on the right track, then the correct analysis for at least some of these inversion errors is that children raise the auxiliary past the subject, that is the children have a non-target grammar of [+TC].[5]

In the discussion above, I have presented some evidence for the existence of contrasting error patterns from Swedish children involving non-target [–TC], as well as errors from English children involving non-target [+TC]. In the following section I review why this is problematic for a deterministic learner and sketch how these errors might arise under the learning model I have proposed.

2.3 Toward accounting for the Swedish and English errors

Let us take stock of the errors we have seen in the last two sections. We saw evidence from V3 child productions in Swedish that the learner sometimes adopts a non-target [–TC] grammar. This contrasts with the adult grammar, which is [+TC]. We saw evidence from subject-aux inversion errors in English children that learners sometimes have a non-target [+TC] grammar, whereas the target grammar for English is [–TC]. These patterns of development are what I have called contrasting errors.

Contrasting errors are not predicted under a deterministic learning model that has a universal set of defaults. Recall that such a model learns only from unambiguous evidence in an attempt to avoid mis-setting parameters, and thus errors must be the result of the default value. No matter what the default is, though, only one set of errors (either those in Swedish or English) is predicted under this approach. For example, suppose that the default value is [–TC]. Give this initial state, then as the Swedish learner moves toward a [+TC] grammar, V3 errors in Swedish are possible. However, we do not expect subject-aux inversion errors for English learners, because there is no unambiguous evidence for English being an across-the-board [+TC] language. Such evidence would be, for example, unambiguous input for T-to-C movement in declaratives. The converse holds if the default is [+TC]. With such an initial state, as the English learner moves toward a [–TC] grammar, subject-aux inversion errors in non-interrogatives are possible. However, we do not expect to see Swedish errors with the finite verb in T. It is possible that Swedish errors could involve the verb remaining in-situ if the default is [–VT]. However, if we see productions in which the verb is raising to T, it must continue on to C because of the [+TC] default.

[5] If children are sometimes adopting a non-target [+TC] grammar, one might wonder whether we also expect to see aux-initial errors in declaratives of the form #[AuxSV(O)]. In note 4 I discussed Pierce's (1992) observation that errors of this sort are indeed attested. I would like to point out, though, that these are not necessarily robustly expected. Crucially, these errors depend on the absence of a some constituent in SpecCP. If children have learned that verb movement to C often or always co-occurs with some other constituent being in SpecCP (perhaps sometimes a phonologically null question operator), then an utterance with T-to-C movement might look like an ordinary string of the form [SAuxV(O)]. Formalizing how the model learns what is in SpecCP is an issue I have not looked at in this thesis, and is a topic that merits further investigation.

I conclude that a deterministic model that learns only from unambiguous evidence cannot account for the full range of child errors that we see in language acquisition. A deterministic learner is designed to avoid errors. I have discussed how such a learner in general needs defaults because of cases where there is insufficient unambiguous evidence. Defaults also provide a deterministic learner with a way of accounting for learner errors. Nevertheless, defaults predict only a subset of the errors that are actually attested. The remaining errors are left unaccounted for with such a model.

In contrast, a probabilistic model that learns from ambiguous evidence provides us with a plausible way of making sense of these contrasting errors in Swedish and English. Here I will simply sketch what the relation between ambiguous evidence and these errors is. Formally modeling this relation awaits future research. In sketching a proposal for the Swedish and English errors, what I will do is to point out that in both languages the input is highly ambiguous between [±TC].

I begin again with Swedish. My account of Swedish parallels that of Swiss German in Chapter 3. V3 errors were also attested in Swiss German, albeit much more rarely, and these were attributed to a non-target [–TC] grammar that was also [+VT]. With Swiss German we saw that there was robust unambiguous evidence for raising the verb (around one-third of the input). This evidence was thus unambiguous for [+VT]. Extrapolating from the corpus analysis in Lightfoot (1997: 265), Swedish and German have a comparable amount of unambiguous input for verb raising.

However, all input tokens without embedded clauses in Swiss German were ambiguous as to where the verb actually raised. The verb could have raised to either T or C. Further, many embedded clauses did not resolve this ambiguity. This ambiguity is exactly what we have seen in matrix clauses in Swedish, such as (1), repeated below.

(1) a. Författaren skrev inte någen bok i år.
author.the wrote not any book in year
'The author didn't write any book this year.' (Waldmann 2011: 332)

b. [$_{CP}$ C [$_{TP}$ V_{fin}+T [Neg … ~~V_{fin}~~ …]]]

c. [$_{CP}$ V_{fin}+T+C [$_{TP}$ Neg [$_{TP}$ ~~V_{fin}+T~~ [… ~~V_{fin}~~ …]]]]

Additionally, the embedded evidence in (2), again repeated here, does not provide unambiguous evidence for [+TC], although as we saw above, a consistent analysis of (1) and (2) is only possible with a [+TC] grammar. On its own though, (2) is ambiguous: even if the grammar is [+TC], the verb will still appear in T because of the presence of the overt complementizer.

(2) a. Vi frågade [om författaren inte skrev någen bok i år].
we asked if author.the not wrote any book in year
'We asked if the author didn't write any book this year.' (Waldmann 2011: 332)

b. …[$_{CP}$ if [$_{TP}$ author [$_{TP}$ Neg [$_{TP}$ wrote+T [… ~~wrote~~ …]]]]]

Crucially, this ambiguity for T-to-C movement can be found in the majority of the input. For Swiss German, I estimated that around 87% of the input contained no embedded clause and thus was ambiguous for [±TC]. Further some of the input with embedded clauses is ambiguous for T-to-C movement as well. This pervasive ambiguity in the input played a key role in allowing the learner to temporarily maintain [–TC] as a viable hypothesis, as we saw in the modeling results for Swiss German in Chapter 3. Swedish has a comparable rate of embedding. Waldmann's (2014: 65-66) corpus analysis found that in child directed Swedish, around 12% of utterances with a verb contained an embedded clause. And again, some of the input with embedded clauses is ambiguous for T-to-C movement. Given this ambiguity, we also expect to see some Swedish learners temporarily adopt a [–TC] grammar.[6] Thus under the learning model, it is expected that ambiguous evidence plays a role in Swedish learners adopting a non-target [–TC] grammar. Coupled with the robust evidence for [+VT], then similar to what we saw for Swiss German, we expect the model to pass through a non-target state in which V3 errors (at least for some children) are possible. Further, the learner can be influenced by this ambiguous evidence to adopt the non-target grammar if there is no strong bias for the target values of [+TC], as is the case if the learner has equal priors and thus no bias. The proposed account of Swedish V3 errors, then, follows the familiar contours of the model's temporary indeterminancy in the face of ambiguous input.[7]

Ambiguous evidence can also push the learner toward a [+TC] grammar in English. First we can observe that many basic utterances that the learner hears – for example, utterances of the form SAuxV(O) – are ambiguous with respect to the position of the auxiliary. This is the same basic ambiguity regarding verb placement that we have now seen in Swiss German and Swedish. Is the auxiliary in T or in C? The fact that there is such widespread ambiguity means that the learner cannot readily reject [+TC] as a viable hypothesis, although unambiguous input, such as (6), repeated here, can push the learner toward a [–TC] grammar.

(6) a. Mary probably has (not) swum today.
b. I know [which homework assignment Mary has not done yet]!

[6] Waldmann's (2014: 53, 63) analysis of corpus directed Swedish found that of utterances containing verbs, less than 1% of these utterances contained embedded clauses with negation or a sentence-medial adverb (both of which pattern together as regards word order). I have suggested that this sort of input is crucial in learning a [+TC] grammar in Swedish. Given that this frequency is much less than other input frequencies in the simulations run in the case studies, it is reasonable to wonder how well the model would learn a [+TC] grammar if relying on this kind of embedded clause input to do so. This remains a question to be investigated in future simulations.

[7] Thus I agree with Waldmann (2011) in analyzing the Swedish child errors as having the verb in T (and not in C), but I do not follow Waldmann's claim that these errors result from some learning strategy involving economy of movement. According to an account of these errors based on economy, learners initially prefer the analysis with less movement to T over the analysis involving more movement to C.

I do not know how frequently attested input such as (6) in child directed speech. Thus it is unclear how strongly unambiguous evidence pushes the learner away from [+TC]. However there is robust ambiguous evidence that can mislead the learner into thinking that English is a [+TC] language. These are the cases of subject-aux inversion that we see in interrogatives.[8] Recall that questions with subject-aux inversion are compatible with either a more general [+TC] setting, or with a [–TC] grammar that is set to [+TC-Q]. In fact, the majority of grammars compatible with questions are [+TC] – that is, the number of input-compatible [+TC, ±TC-Q] grammars is great than the number of those that are [–TC, +TC-Q] because, all else being equal, the value of [±TC-Q] need not be fixed in the case of the former. Questions are certainly widely found in child directed speech, and the subject-aux inversion found in this input can push the learner toward temporarily reinforcing a non-target [+TC] value.[9] Thus the children's errors resemble an important source of ambiguous input that they encounter, and the nature of the hypothesis space concerning this input favors some other non-target grammar. Ultimately, we expect the target values of [–TC, +TC-Q] to emerge as the grammar of best fit, as those are the values that are consistent will all the input. During the acquisition process, though, so long as [+TC] has been sufficiently reinforced, there will be a delay in the learner's rejecting it, and subject-aux inversion errors are possible, including in non-interrogatives. According to this proposal, then, English subject-aux inversion errors can also be attributed to the ambiguity in the input, and in particular, common input that favors a non-target parameter setting. Again it is the pervasiveness of this ambiguity that plays a role in the learner's temporarily adopting a non-target hypothesis, which will later be rejected. And again the non-target [+TC] grammar can be adopted if there is not too strong of a bias for [–TC]; this is in line with the null hypothesis, which has equal priors.

Putting the pieces together, we can say the following about the contrasting error patterns and bias. Given Swedish, the learner cannot be overly biased toward [+TC], and given English, the learner cannot overly biased toward [–TC]. This is captured with equal priors. With equal priors the learner has the greatest likelihood of choosing either parameter value. Thus, with these non-biased priors, the learner has the greatest likelihood of making errors that reflect a different non-target parameter value depending on the language. Thus a model with non-biased priors is likely to have the greatest success at accounting for contrasting error patterns cross-linguistically.

---

[8] Stromswold (1990: 213) also suggests that hearing input involving subject-aux inversion in matrix *wh*-questions plays a role in the embedded clause inversion errors, but she does not attribute these errors to some more general property of the input, namely that the input contains inversion.

[9] For example, Yang (2002: 43) estimates based on corpus analysis that 30% of sentences in child directed English contain *wh*-questions, although only questions in which the *wh*-word is a non-subject will contain subject-aux inversion. Further, if we extrapolate from Westergaard's (2006) corpus study of child-directed Norwegian, we can also estimate that there is a high proportion of yes/no polar questions, which also contain subject-aux inversion. Westergaard found that polar questions also occur in around 30% of the input sentences.

In summary, the contrasting error patterns in Swedish and English pose a problem for a deterministic learner in general, and in particular one that is equipped with biases in the form of a universal set of defaults. Such a model does not learn from ambiguous evidence and can account for the errors in at most one of the two languages. In contrast, a model that can exploit the widespread ambiguity of the input provides us with a plausible account of the cross-linguistic errors. This is possible if the learner is not overly biased toward one particular parameter value over another, and this is captured well by the null hypothesis of non-biased priors. Errors are simply the result of sufficiently reinforcing non-target parameter values in the face of a substantial amount of ambiguous evidence. And this is especially true when a preponderance of the input-compatible grammars for some common input favor a non-target parameter value. This discussion of learning errors thus provides support for the approach I have been taking, according to which I have been assuming that in general the learner's priors are non-biased.

3. Constraining the model: A new proposal

In this section I consider how a learning bias could help constrain the learning model with regard to the ease of learning certain grammars as end-states. This kind of a bias would help the model to account for certain language universals according to which certain grammars are not attested.[10] I will focus here on a single such universal, the Final-over-Final Constraint, or FOFC, (Biberauer et al. 2014), which was discussed in some detail in Chapter 4. In its current form, the model cannot account for this universal. In an attempt to account for this universal I will propose a universal (violable) constraint in the grammar concerning relations between parameter values. This constraint is linked to a bias in how parameter values are reinforced. To the extent that this proposal can account for the universal, then it also provides an illustration of how to implement something like a principle of the Principles and Parameters framework in the model, as both the constraint and the bias would be invariant across all children learning language.[11]

Recall that according to FOFC, a language-type that is unattested is one in which, within certain domains such as the extended projection of the verb, a head-final phrase takes a head-initial phrase as its complement. For example, a language in which a head-final TP takes a head-initial VP is claimed to be unattested. As was discussed in Chapter 4, nothing constrains the model from learning a FOFC-violating grammar as an end-state.

[10] The kind of bias explored here could in principle be used to model cross-linguistic tendencies as well, such as for, example the correlation between VP-headedness and PP-headedness discussed in Dryer (1991).

[11] This conception of a principle as involving a violable constraint is perhaps different from, but I believe very much in the spirit of, how principles are typically discussed in the Principles and Parameters framework. Thus what I have in mind is not some property that categorically rules out certain structures (as indeed Biberauer et al. (2014) seem to have in mind in accounting for FOFC), but rather some property of the grammar that is invariant across learners. This also relates to the empirical question I raised in Chapter 4 (note 14) about whether a FOFC-violating grammar is unlearnable. The constraint and bias I propose mitigate against learning a FOFC-violating grammar, but do not completely rule out learning such a grammar on their own (cf. note 12 for discussion of this and why it need not be problematic).

All parameter values are reinforced equally according to the shape of the input, and if [T-fin] and [V-init] are both a good fit to the input, then the current implementation of the model trivially learns a [T-fin, V-init] grammar.

In attempting to constrain the model there are two goals. First, the goal is to capture the universal, in the case at hand to not arrive at an end-state that is FOFC-violating. Second, we want to maintain the empirical coverage of the model's current learning results. In Chapter 4 we saw that the challenge was to rule out a FOFC-violating end-state while still accounting for the variability we see in Korean. My proposal is that a learning bias might be able to accomplish these goals. In short, the proposal is that when a FOFC-violating grammar is sampled that fits the input, some values are only partially reinforced. The proposal here is speculative, and should be tested to see how well it fares. Nevertheless, unlike the proposal to constrain the model in Chapter 4, a bias of partial reinforcement appears better suited to meeting the goals I have presented. Let us now consider in more detail what this bias looks like.

First, we can observe that in all simulations of the model up to this point, reinforcement has been non-biased. Thus far, whenever a set of parameter values is updated, all the values are updated equally. As an example, consider TP-headedness and VP-headedness. If the model has sampled the FOFC-violating set of values [T-fin, V-init], and if these choices are to be reinforced because they fit the input datum, then both values are reinforced by the same amount. In terms of these values' pseudo-count totals, we can say that each value's total increases equally, say by an increment of 1. I have kept constant throughout the simulations the increment by which values are reinforced for each token of input. Let us call this increment in cases of non-biased reinforcement the norm.

In implementing the learning bias, the basic idea is that certain parameter values are reinforced less (i.e. by a smaller increment) than other values (cf. Wilson 2006 and Do 2013 for implementations of this idea). In particular, to capture FOFC we can suppose that learning one parameter value can bias the learner against learning some other parameter value. FOFC concerns the relation between the values of two parameters, namely a head-final value for some phrase and a head-initial value for that phrase's complement. Suppose there is a universal constraint in the grammar against such a relation. Further, suppose this constraint is violable in that the model can still sample and reinforce values that instantiate such a relation, for example [T-fin, V-init]. The effect of the constraint is that if it is violated, then the bias kicks in. According to the bias, some value or values will be reinforced less given such a sample. Minimally, one of the two values in the relevant relation that is constraint-violating must be a non-target one. The bias works to reinforce one of those values by some increment smaller than the norm. The other value will still be reinforced by an increment equal to the norm. It does not matter which value is reinforced less, so long as the bias is consistent against the same value (either initial or final) across instances of a constraint violation. For example,

assume there is a bias against V-initial. Then whenever the model reinforces a sample that violates the constraint with [T-fin, V-init], it is V-initial that will be reinforced by some increment smaller than the norm. How small is this increment? The size of this increment can be proportional to the strength of the weight for the other parameter value in the constraint-violating relation, here T-final. The stronger the weight for T-final is, the smaller the increment V-initial is reinforced by; the weaker the weight for T-final is, the closer to the norm is the increment that V-initial is reinforced by.

The effect of this bias is to decrease the force of the evidence when the learner is considering a grammar that violates the constraint. Further, the strength of this evidence is also conditioned by the learner's other expectations. If the learner has strong beliefs concerning the grammar being T-final (i.e. a strong weight for T-final), perhaps because of a sufficient amount of unambiguous evidence for T-final, then the learner is much less likely to expect a V-initial grammar. In such a scenario, a V-initial grammar would be FOFC-violating, and that is the sort of grammar that the bias mitigates against adopting. Conversely, if the learner has weaker expectations for T-final, then V-initial would be reinforced by an increment that is closer to the norm. This is desirable in a scenario where the target grammar is [T-init, V-init]. With only a weak bias against V-initial (because of the weak expectations for T-final), the learner will be deterred less from arriving at the intended V-initial grammar.

The learning model as augmented with this bias now has the potential to make learning a FOFC-violating grammar very unlikely as an end-state, thereby capturing the language universal. Because of the bias, the model will now systematically reinforce FOFC-compatible sets of parameter values more strongly than FOFC-violating sets of values. In general, this will push the learner more and more rapidly toward a FOFC-compatible grammar. This has the effect of making it hard to learn a FOFC-violating grammar as an end-state.

How does the model fare when confronted with input that is only compatible with a FOFC-violating grammar? As I discussed in Chapter 4, if it is indeed difficult to learn a FOFC-violating grammar as an end-state, then even when encountering this 'FOFC-violating input' the model might still end up with a FOFC-compatible grammar. Here I will sketch out one possibility of such a hypothetical scenario involving FOFC-violating input that shows that the model is still pushed toward a FOFC-compatible grammar. I leave for future research a fuller investigation of what the limitations of the model are in all such hypothetical scenarios.

Suppose the learner encounters some input that is compatible only with a FOFC-violating [T-fin, V-init] grammar. Following the discussion above, the learner must reinforce both values, but will reinforce V-initial more weakly. The more evidence of this sort the learner hears, the smaller the increment it will reinforce V-initial by. Given sufficient evidence of this sort, the learner will have a parameter setting of [T-fin], but because of the bias, the weight for V-initial will conceivably have reached some limit

short of that of parameter setting, and beyond which further reinforcement for V-initial is negligible. So far in this hypothetical scenario, V-initial seems to be doing well (even though it is only partially reinforced). Still, given a parameter setting of [T-fin] and the diminishing increments by which V-initial is reinforced, the learner could not sufficiently reinforce V-initial to arrive at a [V-init] parameter setting. Moreover, consider what happens in this scenario when the learner encounters a second type of input that is ambiguous for VP-headedness. For simplicity, let us suppose this input is unambiguous evidence for T-final, although this point is not crucial so long as parameter interaction does not favor V-initial with this second type of input (i.e. the grammars compatible with this input are distributed equally between V-initial and V-final). Given this ambiguous input, the learner could reinforce either V-initial or V-final, but the learner will systematically reinforce V-final more strongly. The more of this ambiguous input the learner encounters, the more the balance of evidence shifts in favor of a [V-fin] grammar. Given a sufficient proportion of this ambiguous evidence in the input corpus, the learner will eventually be pushed toward a FOFC-compatible [T-fin, V-fin] grammar despite the input corpus having FOFC-violating input.[12] The scenario above has been intended to show some of the promise that the learning bias has in constraining the model from reaching a FOFC-violating end-state. It remains to be seen with further simulations of a wider range of possible input corpora how viable this learning bias is as an approach to capture the FOFC universal.

But what about the second goal? With this bias, would the model still be able to cover the same empirical ground as before? In particular, would the model still be to account for the variability we saw in Korean? Again, the learning model as augmented with this bias has the potential to do so. In Chapter 4, we saw that an earlier attempt to constrain the model failed to capture the variability because the learner never considered

[12] Thus in saying the bias constrains the model from reaching a FOFC-violating end-state, there is the following caveat. The model is still sensitive to what the actual input corpus is. Consider an input corpus that contained only FOFC-violating input all of which was compatible only with a [T-fin, V-init] grammar. As discussed in the main text, given a sufficiently strong learning bias, the expectation is that the model will have strong enough weighting for a [T-fin] parameter setting, but will have insufficient weight on either value to have a parameter setting of [V-init] or [V-fin]. The learner will reach an end-state that is probabilistically V-initial and V-final, albeit one that favors V-initial. Such an end-state is clearly FOFC-violating, at least some of the time. This input corpus differed from the hypothetical input corpus in the text in that it contained no ambiguous input for the parameters under consideration. Crucially, learning a FOFC-compatible grammar was attributed to there being a sufficient amount of ambiguous evidence in the input. In attempting to account for the FOFC universal, is this kind of input sensitivity worrying? Not obviously. Throughout this thesis we have seen that a pervasive property of the languages we have looked at is that for any given parameter they contain input that is ambiguous for that parameter. It seems to me implausible that a natural language input corpus would not contain a substantial amount of input that was similarly parametrically ambiguous. The topic of input-sensitivity surely relates to the question of what other constraints (perhaps extra-linguistic) there might be in determining what a natural language input corpus looks like. Clearly what can be investigated in future work is to what extent the shape of the input corpus plays a role in learning a FOFC-violating end-state. However, without a better understanding of what some of these other constraints are, it seems premature to declare the inadequacy of the learning bias I have proposed on the basis of this kind of input sensitivity.

samples of parameter values that were FOFC-violating. This constraint was implemented with a filter on what kinds of samples could have their values be reinforced. The result of this FOFC-filter was to favor a non-verb raising [–V-*v*] grammar by a two-to-one margin among the sampled grammars that passed through the filter. This is illustrated in (14) and (15), repeated from Chapter 4.

(14) *Grammars compatible with SOV input from the 5-parameter model, Chapter 4*

a. FOFC-compatible	b. FOFC-violating
1. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]	1. [+V-*v*, +*v*-T, T-fin, *v*-in, V-in]
	2. [+V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
	3. [+V-*v*, +*v*-T, T-fin, *v*-fin, V-in]
2. [+V-*v*, –*v*-T, T-in, *v*-fin, V-fin]	4. [+V-*v*, –*v*-T, T-in, *v*-fin, V-in]
3. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]	5. [+V-*v*, –*v*-T, T-fin, *v*-fin, V-in]
4. [–V-*v*, +*v*-T, T-in, *v*-in, V-fin]	6. [–V-*v*, +*v*-T, T-fin, *v*-in, V-fin]
5. [–V-*v*, +*v*-T, T-in, *v*-fin, V-fin]	
6. [–V-*v*, +*v*-T, T-fin, *v*-fin, V-fin]	
7. [–V-*v*, –*v*-T, T-in, *v*-in, V-fin]	7. [–V-*v*, –*v*-T, T-fin, *v*-in, V-fin]
8. [–V-*v*, –*v*-T, T-in, *v*-fin, V-fin]	
9. [–V-*v*, –*v*-T, T-fin, *v*-fin, V-fin]	

(15) *Partial summary of grammars in (14a)*
a. +V-*v* grammars: 33.33%
b. –V-*v* grammars: 66.66%

The situation above contrasted with the 50-50 distribution when there was no FOFC-filter between raising and non-raising grammars (among input-compatible grammars). With a FOFC-filter we expected the model to learn only a non-verb raising grammar, whereas when there was no filter, we saw that the model could learn either grammar.

The learning bias I have proposed does not obviously have the disastrous effect of constraining the model with the FOFC-filter. This is in part because the constraint I have proposed is violable. During the learning process, the learner can consider and reinforce both FOFC-violating and FOFC-compatible sets of parameter values. For example with regard to the grammars in (14), the learner can reinforce both values of [*v*-fin, V-init] in a particular sample. Further, the varying strength of the learning bias (according to the learner's varying expectations of different parameter values) also plays a role. According to the bias, early in the learning process values are reinforced more equally. Along the lines of the discussion above, the strength of the bias against V-initial in a [*v*-fin, V-init] sample is proportional to the strength of the weight for *v*-final. Given non-biased priors, on average the learner will not have a strong weight for *v*-final early in development. This means that even when reinforcing a [*v*-fin, V-init] sample, V-initial will be reinforced close to the norm. Recall that there are more input-compatible verb raising grammars that

are FOFC-violating than input-compatible verb raising grammars that are FOFC-compatible. The net effect of the learning bias is that early in the learning process we expect to see the values in sampled grammars with verb raising to be reinforced almost as strongly as those in sampled grammars without verb raising. Given a learner that is not especially tentative, and given the randomness of the sampling procedure, then this discussion suggests that it is possible for the learner to favor a [+V-*v*] verb raising grammar by a meaningful margin early in the learning process. If this edge to a verb raising grammar is sufficiently reinforced (as can happen given the randomness of the sampling procedure), then it is plausible the model could learn a verb-raising grammar. We saw in Korean that parameter interaction will push the learner toward head-final parameter values. As the weights for these head-final values increase, the strength of the bias increases, which in turn means that, all else being equal, a verb raising grammar will be reinforced less strongly than a non-verb raising grammar. Thus if a [+V-*v*] verb raising grammar does get somewhat favored earlier in the learning process, in order to learn a verb raising grammar, the value for [+V-*v*] would presumably need to be sufficiently reinforced before the weight for [*v*-fin] gets too strong (the strength of which increases the strength of the bias). If this discussion is on the right track, then constraining the model with the learning bias would not prevent it from variably learning a grammar with or without verb raising. This can be tested in future simulations of the learning model.

In this section I have attempted to show how constraining the model with a learning bias has the potential to account for a putative language universal, as well as capturing the variability in learning verb movement that we saw in Korean in Chapter 4. The learning bias involved partially reinforcing some parameter values, such that learning certain parameters made it intrinsically harder to learn certain other parameter values. This has the effect of constraining what a likely end-state for the model is. Again, further simulations of the model are necessary to evaluate the success of this learning bias. Nevertheless, it offers a not unpromising way of moving forward to address some of the potential challenges in constraining the learning model.

4. Summary

In this chapter, I returned to several issues from earlier in this thesis in the context of learning biases. First, I showed that a deterministic model that learns only from unambiguous evidence cannot account for the contrasting error patterns we see with Swedish and English children. In particular, this was true in the scenario where a deterministic learner was augmented with defaults, which can be represented by biased priors. Then I sketched how a probabilistic model that learns from ambiguous evidence with non-biased priors might account for these errors. Second, I discussed how a learning bias might help the model to account for a language universal, namely the Final-over-Final Constraint (FOFC). The implementation of this bias took the form of partially reinforcing some parameter values in proportion to the reinforcement of other parameter

values. I also discussed how this implementation has the potential to still be able to model variability across speakers. Both of these topics raised questions for future work. On the one hand, we would like to test the learning model here to see if it is able to account for the learning errors in Swedish and English. And on the other hand, we would like to see exactly how successful the learning bias is in accounting for FOFC, as well as modeling variability across speakers.

Chapter 7

Final Summary

In this thesis, I have presented a probabilistic model for parameter setting that learns from ambiguous evidence. I discussed the challenges that ambiguous evidence presents from a learnability perspective. Then I ran simulations of the model to give proof-of-concept illustrations of how the model can address these challenges. In particular, ambiguous evidence poses a learnability difficulty when there is no unambiguous evidence for some set of parameters. Examples of this included the subset case of zero-derived causatives in English and phrase structure parameters in Korean. Nevertheless, the model is successful at parameter setting by learning the grammar of best fit to the input. One way this can be done is by learning via implicit negative evidence. Further, a novel insight was to show how such a grammar can also be learned via parameter interaction.

The investigation of parameter interaction also showed how we can model two empirical phenomena in language acquisition. Sometimes multiple grammars are highly compatible with the learner's evidence. In such a scenario, the model could learn either grammar. An application of the model, then, was to show how it could be used to account for grammatical variability across learners, as in the case of verb movement in Korean. A further implication of parameter interaction was to show that during the course of learning, the grammar of best fit is sometimes a non-target grammar. Thus a further application of the model was to show how it can be used to account for learner errors, as in the case of verb placement in embedded clauses in Swiss German.

Having seen how the model works in the case studies mentioned above, we can now ask a more general question about how to evaluate the model's overall success. In a sense, this is a question about how well the model fares with future case studies (including enriched versions of the ones presented in this thesis). The implementations of the model I have presented involve many simplifications. Indeed, that is the very nature of what a model is, and at some point of future inquiry we are likely to see some aspect of the model that is wrong and needs changing. The hope is that the model has provided some insight into the learning process. Insight into how a lack of unambiguous evidence need not be an impediment to language acquisition. Insight into how grammatical variability arises. Insight into the nature of children's acquisition errors. If the model has provided some insight on these issues, as I believe it has, then as we look beyond the case studies covered here, we would like to see what the limits of the model are in accounting for a broader range of the phenomena I have touched on.

Chief among these limits is whether the model can account for the full typology of learner errors during the course of language acquisition. According to the model, the kinds of errors we expect to see can be attributed to the ambiguity of the input, especially input for which a preponderance of the compatible grammars favors a non-target

hypothesis. Further, the prevalence of these errors (both within speakers and across a population), as well as the persistence of these errors over time, is to be attributed to factors such as the frequency of ambiguous input, the order in which the learner hears the input, and how strongly such input favors a non-target hypothesis. Thus the model clearly has something to say about a typology of errors, but how does it actually fare in accounting for such a typology? This is an important question, but attempting to answer it now seems to me to be rather premature for the simple reason that we do not know enough about what children's errors are. We have already seen an example of this with regard to Swiss German in the discussion of embedded clause verb placement errors in other varieties of German. In a large population of children learning a variety of German dialects, how many do we expect to produce these errors and at what kind of developmental trajectory? The answer is not currently known. Not enough children have been studied and certainly not enough embedded clause productions of those that have. More generally, some errors, especially those related to parameters for basic word order, might appear only fleetingly, and only with some children. These errors could easily fail to be documented in a small population of children and in corpora that are only samplings of a child's spontaneous speech. In other words, the typology of errors is still emerging. The best approach I see, moving forward, is an empirically based one, in which we gather more data on which to test the model.

References

Abe, Jun. 1993. *Binding Conditions and Scrambling without A/A' Distinction*. PhD dissertation, The University of Cennecticut.

Anderssen, Merete, Kristine Bentzen, Yulia Rodina, and Marit Westergaard. 2011. "The Acquisition of Apparent Optionality: Word order in subject and object shift constructions in Norwegian", in Merete Anderssen, Kristine Bentzen, and Marit Westergaard (eds.), *Variation in the Input: Studies in theoretical psycholinguistics*. Netherlands: Springer, pp. 241-270.

Atkinson, Martin. 2001. "Learnability and the Acquisition of Syntax", in Stefano Bartolo (ed.), *Language Acquisition and Learnability*. Cambridge: Cambridge University Press, pp. 15-80.

Bach, Emmon. 1962. "The Order of Elements in a Transformational Grammar of German", *Language* 38 (3): 263-269.

Baker, Mark. 2001. *The Atoms of Language: The mind's hidden rules of grammar*. New York: Basic Books.

Berwick, Robert. C. 1986. "Learning from Positive-Only Examples: The subset principle and three case studies", in Ryszard. S. Michalski, Jaime. C. Carbonell, and Tom. M. Mitchell (eds.), *Machine learning: An artificial intelligence approach, Volume 2*. San Mateo, CA: Morgan Kaufmann, pp. 625-645.

Biberauer, Theresa, Anders Holmberg, and Ian Roberts. 2014. "A Syntactic Universal and Its Consequences", *Linguistic Inquiry* 45 (2): 169-225.

Blanco, Mercedes Tubino. 2010. *Contrasting Causatives: A Minimalist approach*. PhD dissertation, The University of Arizona.

Bobaljik, Jonathan David. 2000. *The Rich Agreement Hypothesis in Review*. Ms., McGill University, Montréal.

Bobaljik, Jonathan David. 2001. "The Implication of Rich Agreement: Why morphology doesn't drive syntax", in Karin Megerdoomian and Leora Bar-el (eds.), *Proceedings of the 20th West Coast Conference on Formal Linguistics (WCCFL 20)*. Somerville, MA: Cascadilla Press, pp. 82-95.

Braine, Martin D. S. 1971. "On Two Types of Models of the Internalization of Grammars", in Daniel Slobin (ed.), *The Ontogenesis of Grammar: A theoretical symposium*. New York: Academic Press, pp. 153-186.

Boersma, Paul. 1997. "How We Learn Variation, Optionality, and Probability", in Rob J. J. H. van Son (ed.), *Proceedings of the Institute of Phonetic Sciences 21*. Amsterdam: The University of Amsterdam, pp. 43-58.

Borer, Hagit. 2005. *Structuring Sense, Volumes I and II*. Oxford: Oxford University Press.

Bowerman, Melissa. 1988. "The 'No Negative Evidence' Problem: How do children avoid constructing an overly general grammar?", in J. A. Hawkins (ed.), *Explaining Language Universals*, Oxford: Basil Blackwell, pp. 73-101.

Cho, Sook Whan. 1981. *The Acquisition of Word Order in Korean*. MA Thesis, The University of Calgary.

Cho, Young-Mee Yu and Ki-Sun Hong. 1988. "Evidence for the VP Constituent from Child Korean", *Papers and Reports on Child Language Development 27*. Stanford: Department of Linguistics, Stanford University.

Cho, Sae-Youn. 1993. "Auxiliary Verb Constructions in Korean", *Studies in the Linguistic Sciences* 23 (2): 1-24.

Cho, Young-Mee Yu and Peter Sells. 1995. "A Lexical Account of Inflectional Suffixes in Korean", *Journal of East Asian Linguistics* 4 (2): 119-174.

Choe, Hyon Sook. 1987. "Successive-cyclic Rightward Movement in Korean", in Susumu Kuno et al. (eds.), *Harvard Studies in Korean Linguistics 2: Proceedings of the 1987 Harvard Workshop on Korean Linguistics*. Seoul: Hanshin Publishing Company, pp. 40-56.

Choi, Young-Sik. 1999. "Negation, its scope and NPI Licensing in Korean", in Rebecca Daly and Anastasia Riehl (eds.), *ESCOL '99*. Ithaca: Cornell University, CLC Publications, pp. 25-36.

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

Chomsky, Noam. 1980. *Rules and Representations*. Oxford: Basil Blackwell.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam. 1986. *Knowledge of Language: Its nature, origin, and use*. New York: Praeger.

Chung, Daeho. 2005. "What Does Bare *–ko* Coordination Say about Post-verbal Morphology in Korean?" *Lingua* 115 (4): 549-568.

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads: A cross-linguistic perspective*. Oxford: Oxford University Press.

Cinque, Guglielmo. 2006. *Restructuring and Functional Heads: The cartography of Syntactic Structures, Volume 4*. Oxford: Oxford University Press.

Clahsen, Harald. 1982. *Spracherwerb in der Kindheit. Eine untersuchung zur entwicklung der syntax bei kleinkindern*. Tübingen: Gunter Narr Verlag.

Clahsen, Harald and Klaus-Dirk Smolka. 1986. "Psycholinguistic Evidence and the V-Second in German", in Hubert Haider and Martin Prinzhorn (eds.), *Verb Second Phenomena in Germanic Languages*. Dordrecht: Foris, pp. 137-167.

Clark, Robin. 1989. "On the Relationship Between the Input Data and Parameter Setting" in Juli Carter and Rose-Mari Déchaine (eds.), *Proceedings of the Northeast Linguistic Society (NELS) 19*. Amherst: Graduate Linguistic Student Association, The University of Massachusetts, pp. 48-62.

Clark, Robin. 1992. “The Selection of Syntactic Knowledge”, *Language Acquisition* 2 (2): 85-149.

Clark, Robin and Ian Roberts. 1993. “A Computational Model of Language Learnability and Language Change” *Linguistic Inquiry* 24 (2): 299-345.

Crain, Stephen and Rosalind Thornton. 1998. *Investigations into Universal Grammar: A guide to experiments in the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

Davies, Mark. 2008-2015. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

den Besten, Hans. 1977/1989. *Studies in West Germanic Syntax*. Amsterdam: Rodopi.

den Dikken, Marcel. 2006. *A Reappraisal of vP Being Phasal*. Ms., CUNY Graduate Center, New York City.

Déprez, Viviane and Amy Pierce. 1993. “Negation and Functional Projections in Early Grammar” *Linguistic Inquiry* 24 (1): 25-67.

Diesing, Molly 1992. *Indefinites*. Cambridge, MA: MIT Press.

Do, Young Ah. 2013. *Biased Learning of Phonological Alternations*. Ph.D. dissertation, Massachusetts Institute of Technology.

Dowty, David R. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Dryer, Matthew S. 1991. “SVO Languages and the OV/VO Typology”, *Journal of Linguistics* 27 (2): 443-482.

Embick, David and Ralf Noyer. 2001. “Movement Operations after Syntax”, *Linguistic Inquiry* 32 (4): 555-595.

Ernst, Thomas. 2002. *The Syntax of Adjuncts*. Cambridge: Cambridge University Press.

Fodor, Janet Dean. 1998. “Unambiguous Triggers”, *Linguistic Inquiry* 29 (1): 1-36.

Fodor, Jerry. 1970. “Three Reasons for Not Deriving ‘kill’ from ‘cause to die’”, *Linguistic Inquiry* 1 (4): 429-438.

Folli, Raffaella and Heidi Harley. 2005. “Flavors of *v*: Consuming results in Italian and English”, in Roumyana Slabakova and Paula Kempchinsky (eds.), *Aspectual Inquiries*. Dordrecht: Kluwer, pp. 95-120.

Frank, Robert and Shyam Kapur. 1996. “On the Use of Triggers in Parameter Setting”, *Linguistic Inquiry* 27 (4): 623-660.

Fukuda, Shin and Soonja Choi. 2009. “The Acquisition of Transitivity in Japanese and Korean Children”, in Shoichi Iwasaki, Hajime Hoji, Patricia Clancy, Sung-Och Sohn (eds.), *Japanese and Korean Linguistics 17*. Stanford: CSLI Publications, pp. 613-624.

Gawlitzek-Maiwald, Ira, Rosemarie Tracy, and Agnes Fritzenschaft. 1992. “Language Acquisition and Competing Linguistic Representations: The child as arbiter”, in Jürgen M. Meisel (ed.), *The Acquisition of Verb Placement: Functional categories and V2 phenomena in language acquisition*. Dordrecht: Kluwer, pp. 139-179.

Gibson, Edward and Ken Wexler. 1994. “Triggers”, *Linguistic Inquiry* 25 (3): 407-454.

Gold, E. Mark 1967. “Language Identification in the Limit”, *Information and Control* 10 (5): 447-474.

Goodman, Noah, Vikash Mansinghka, Daiel Roy, Keith Bonawitz, and Joshua Tenenbaum. 2008. “Church: A language for generative models”, in (eds. David McAllester and Petri Myllymaki) *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI-08)*. Corvallis, Oregon: AUAI Press, pp. 220-229.

de Haan, Germen and Fred Weerman. 1986. “Finiteness and Verb Fronting in Frisian”, in Hubert Haider and Martin Prinzhorn (eds.), *Verb Second Phenomena in Germanic Languages*. Dordrecht: Foris, pp. 77-110.

Hagstrom, Paul. 1995. *Negation, Focus, and do-support in Korean*. Ms., Massachusetts Institute of Technology, Cambridge, MA.

Hagstrom, Paul. 2002. “Implications of Child Errors for the Syntax of Negation in Korean”, *Journal of East Asian Linguistics* 11 (3): 211-242.

Haider, Hubert. 2010. *The Syntax of German*. Cambridge: Cambridge University Press.

Hale, Ken and Samuel Jay Keyser. 2002. *Prolegomenon to a Theory of Argument Structure*. Cambridge, MA: MIT Press.

Han, Chung-hye, Jeffrey Lidz, and Julien Musolino. 2007. “V-Raising and Grammar Competition in Korean: Evidence from negation and quantifier scope”, *Linguistic Inquiry* 38 (1): 1-47.

Han, Chung-hye, Dennis Ryan Storoshenko, and Yasuko Sakurai. 2008. “An Experimental Investigation into the Placement of the Verb in the Clause Structure of Japanese”, in *Proceedings of the 2007 International Conference on Linguistics in Korea (ICLK-2007)*. Seoul: The Linguistic Society of Korea.

Han, Chung-hye, Jeffrey Lidz, and Dennis Ryan Storoshenko. To appear. “Variation in Negation and Quantifier Scope Judgments in Korean”, in *Proceedings of the 47th Annual Meeting of the Chicago Linguistic Society*. Chicago: The Chicago Linguistic Society.

Heim, Irene and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Holmberg, Anders. In press. “Verb Second” in Tibor Kiss and Artemis Alexiadou (eds.), *Syntax – an International Handbook of Syntactic Research*. Berlin: Walter de Gruyter.

Hsu, Anne S. and Thomas L. Griffiths. 2009. “Differential Use of Implicit Negative Evidence in Generative and Discriminative Language Learning”, in Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta (eds.), *Advances in Neural Information Processing Systems* 22, pp. 754-762.

Hyams, Nina. 1986. *Language Acquisition and the Theory of Parameters*. Dordrecht: Reidel.

Joo, Yanghee. 1989. *A Cross-linguistic Approach to Quantification in Syntax*. Ph.D. dissertation, The University of Wisconsin, Madison.

Josefsson, Gunlög. 1996/2013. "The Acquisition of Object Shift in Swedish Child Language", in Carolyn E. Johnson and John H. V. Gilbert (eds.) *Child Language, Volume 9*. Psychology Press, pp. 153-165.

Kayne, Richard. 1994. *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Kemp, Charles, Amy Perfors, and Joshua Tenenbaum. 2007. "Learning Overhypotheses with Hierarchical Bayesian Models", *Developmental Science* 10 (3): 307-321.

Kim, Young-Joo. 1997. "The Acquisition of Korean", in Dan Isaac Slobin (ed.), *The Crosslinguistic Study of Language Acquisition, Volume 4*. Mahwah: Lawrence Erlbaum.

Kim, Young-Joo. 2000. "Subject/Object Drop in the Acquisition of Korean: A cross-linguistic comparison", *Journal of East Asian Linguistics* 9 (4): 325-351.

Klima, Edward and Ursula Bellugi. 1966. "Syntactic Regularities in the Speech of Children" in John Lyons and Roger Wales (eds.) *Psycholinguistic Papers*. Edinburgh: Edinburgh University Press, pp. 183-208.

Koizumi, Masatoshi. 2000. "String Vacuous Overt Verb-Raising", *Journal of East Asian Linguistics* 9 (3): 227-285.

Koopman, Hilda. 2005. "Korean (and Japanese) Morphology from a Syntactic Perspective", *Linguistic Inquiry* 36 (4): 601-633.

Koster, Jan. 1975. "Dutch as an SOV Language", *Linguistic Analysis* 1: 111-136.

Kratzer, Angelika. 1996. "Severing the external argument from the verb", in Johann Rooryck and Laurie Zaring (eds.), *Phrase Structure and the Lexicon*. Dordrecht: Kluwer, pp. 109-137.

Kratzer, Kai, Joshua T. Berryman, Aaron Taudt, Johannes Zeman, and Alex Arnold. 2014. "The Flexible Rare Event Sampling Harness System (FRESHS)", *Computer Physics Communications* 185 (7): 1875-1885.

Kroch, Anthony. 1989. "Reflexes of Grammar in Patterns of Language Change", *Language Variation and Change* 1 (3): 199-244.

Kuroda, Shige-Yuki. 1970. *Japanese Syntax and Semantics*. Dordrecht: Kluwer.

Laplace, Pierre-Simon. 1825. *Philosophical Essay on Probabilities*. Translated by Andrew I. Dale (1995) from the fifth French edition. Springer: New York.

Lee, Jae Hong. 1993. "Postverbal Adverbs and Verb Movement in Korean", in Patricia Marie Clancy and Hajime Hoji (eds.), *Japanese/Korean Linguistics 2*. Stanford: SLA/CSLI Publications, pp. 429-446.

Lee, Jungmee. 2008. "The Temporal Interpretation of the Korean –ko Construction: Aktionsart and discourse context", in Atle Grønn (ed.), *Proceedings of Sinn und Bedeutung 12*. Oslo, pp. 367-383.

Lee, Jungmee and Judith Tonhauser. 2010. "Temporal Interpretation without Tense: Korean and Japanese coordination constructions", *Journal of Semantics* 27 (3): 307-341.

Legate, Julie Anne. 2003. "Some Interface Propertiess of the Phase", *Linguistic Inquiry* 34 (3): 506-515.

Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity*. Cambridge, MA: MIT Press.

Lightfoot David. 1989. "The Child's Trigger Experience: Degree-0 learnability" *Behavioral and Brain Sciences* 12 (2): 321-334

Lightfoot, David. 1991. *How to Set Parameters: Arguments from language change*. Cambridge, MA: MIT Press.

Lightfoot, David. 1997. "Shifting Triggers and Diachronic Reanalysis", in Ans van Kemenade and Nigel Vincent (eds.), *Parameters of Morphosyntactic Change*. Cambridge: Cambridge University Press, pp. 253-272.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk*, *Third edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Magri, Giorgio. 2013. "HG Has No Computational Advantages Over OT: Toward a new toolkit for computational OT", *Linguistic Inquiry* 44 (4): 569-609.

Marantz, Alec. 1988. "Clitics, Morphological Merger, and the Mapping to Phonological Structure" in Michael Hammond and Michael Noonan (eds.), *Theoretical Morphology: Approaches in Modern Linguistics*. San Diego: Academic Press, pp. 253-270.

Marantz, Alec. 1997. "No Escape from Syntax: Don't attempt morphological analysis in the privacy of your own lexicon", in Alexis Dimitriadis, Laura Siegel, Clarissa Suerk-Clark, and Alexander Williams (eds.), *Pennsylvania Working Papers in Linguistics* 4 (2): 201-225. Philadelphia: The University of Pennsylvania, Penn Linguistics Club.

Meinunger, André. 2007. "About Object *es* in the German *Vorfeld*", *Linguistic Inquiry* 38 (3): 553-563.

Mills, Anne E. 1985. "The Acquisition of German", in Dan Isaac Slobin (ed.), *The Crosslinguistic Study of Language Acquisition, Vo1ume 1: The Data*. Hillsdale: Lawrence Erlbaum, pp. 141-254.

Miyagawa, Shigeru. 2001. "The EPP, Scrambling, and *Wh*-in-Situ", in Michael Kenstowicz (ed.), *Ken Hale: A life in language*. Cambridge, MA: MIT Press, pp. 293-338.

Müller, Natascha. 1993. *Komplexe Sätze: Der erwerb von comp und von wortstellungsmustern bei bilingualen kindern (Französisch/Deutsch)*. Tübingen: Gunter Narr Verlag.

Otani, Kazuyo and John Whitman. 1991. "V-raising and VP-ellipsis", *Linguistics Inquiry* 22 (2): 345-358.

Park, Myung-Kan. 1994. *A Morpho-Syntactic Study of Korean Verbal Inflection*. Ph.D. dissertation, The University of Connecticut.

Parsons, Terence. 1990. *Events in the Semantics of English: A study in subatomic semantics*. Cambridge, MA: MIT Press.

Payne, John, Geoffrey K. Pullum, Barabara C. Scholz, and Eva Berlage. 2013. "Anaphoric *one* and Its Implications", Language 89 (4): 794-829.

Pearl, Lisa. 2007. *Necessary Bias in Natural Language Learning*. PhD dissertation, The University of Maryland.

Pearl, Lisa and Sharon Goldwater. In press. "Statistical Learning, Inductive Bias, and Bayesian Inference in Language Acquisition", in Jeffrey Lidz, William Snyder, and Joe Pater (eds.), *The Oxford Handbook of Developmental Linguistics.* Oxford: Oxford University Press.

Pearl, Lisa and Jeffrey Lidz. 2009. "When Domain-General Learning Fails and When It Succeeds: Identifying the contribution of domain specificity", in *Language Learning and Development* 5 (4): 235-265.

Pearl, Lisa and Benjamin Mis. 2011. "How Far Can Indirect Evidence Take Us?: Anaphoric *one* revisited", in Laura Carlson, Christoph Hoelscher, and Thomas F. Shipley (eds.) *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci 2011)*. Austin, TX: Cognitive Science Society, pp. 879-884.

Pearl, Lisa and Benjamin Mis. In press. "The Role of Indirect Positive Evidence in Syntactic Acquisition", *Language*.

Penner, Zvi. 1990. "On the Acquisition of Verb Placement and Verb Projection Raising in Bernese Swiss German", in Monika Rothweiler (ed.), *Spracherwerv und Grammatik: Linguistische untersuchungen zum erwerb von syntax und morphologie, Linguistische Berichte Sonderheft 3*. Opladen: Westdeutscher Verlag, pp. 166-189.

Penner, Zvi. 1996. *From Empty to Doubly-Filled Complementizers*: *A case study in the acquisition of subordination in Bernese Swiss German*. Fachgruppe Sprachwissenschaft der Universität Konstanz. Arbeitspapier Nr. 77.

Perfors, Amy, Joshua Tenenbaum, and Terry Regier. 2006. "Poverty of the Stimulus?: A rational approach", in Ron Sun and Naomi Miyake (eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*. Mahwah, NJ: Lawrence Erlbaum, pp. 663-668.

Perfors, Amy, Joshua Tenenbaum, and Elizabeth Wonnacott. 2010. "Variability, Negative Evidence, and the Acquisition of Verb Argument Constructions", *Journal of Child Language* 37 (3): 607-642.

Perlmutter, David. 1978. "Impersonal Passives and the Unaccusative Hypothesis", *Proceedings of the Berkeley Linguistics Society 4*. Berkeley, CA: The University of California, pp. 157-189.

Pierce, Amy. 1992. *Language Acquisition and Syntactic Theory: A comparative analysis of French and English child grammars*. Dordrecht: Kluwer.

Pintzuk, Susan. 2002. "Verb-Object Order in Old English: Variation as Grammatical Competition", in David Lightfoot (ed.), *Syntactic Effects of Morphological Change*. Oxford University Press: Oxford, pp. 276-300.

Platzack, Christer. 1986. "Comp, Infl, and Germanic Word Order", in Lars Hellan and Kristi Koch (eds.), *Topics in Scandinavian Syntax*. Dordrecht: Reidel, pp. 185-234.

Pollock, Jean-Yves. 1989. "Verb Movement, Universal Grammar, and the Structure of IP", *Linguistic Inquiry* 20 (3): 365-424.

Pylkkänen, Liina. 2008. *Introducing Arguments*. Cambridge, MA: MIT Press.

Ramchand, Gillian Catriona. 2008. *Verb Meaning and the Lexicon: A first-phase syntax*. Cambridge: Cambridge University Press.

Regier, Terry and Susanne Gahl. 2004. Learning the unlearnable: The role of missing evidence. *Cognition* 93 (2): 147-155.

Rizzi, Luigi. 1993/1994. "Some Notes on Linguistic Theory and Language Development: The case of Root Infinitives", *Language Acquisition* 3 (4): 371-393.

Rizzi, Luigi. 1997. "The Fine Structure of the Left Periphery", in Liliane Haegeman (ed.), *Elements of Grammar*. Dordrecht: Kluwer, pp. 281-337.

Roberts, Ian. 1997. "Directionality and Word Order Change in the History of English", in Ans van Kemenade and Nigel Vincent (eds.) *Parameters of Morphosyntactic Change*. Cambridge: Cambridge University Press, pp. 397-426.

Rosengren, Inger. 2002. "EEP: A syntactic device in the service of semantics" *Studia Linguistica* 56 (2): 145-190.

Ross, John Robert. 1967. *Constraints on Variables in Syntax*. PhD dissertation, Massachusetts Institute of Technology.

Rothweiler, Monika. 1993. *Der Erwerb von Nebensätzen in Deutschen: Eine pilotstudie*. Tübingen: Max Niemeyer Verlag.

Safir, Ken. 1987. "Comments on Wexler and Manzini", in Thomas Roeper and Edwin Williams (eds.), *Parameter Setting*. Dordrecht: Reidel, pp. 77-90.

Sakas, William Gregory. 2003. "A Word-Order Database for Testing Computational Models of Language Acquisition", in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 415-422.

Sakas, William Gregory and Janet Dean Fodor. 2001. "The Structural Triggers Learner", in Stefano Bartolo (ed.), *Language Acquisition and Learnability*. Cambridge: Cambridge University Press, pp. 172-233.

Sakas, William Gregory and Janet Dean Fodor. 2012. "Disambiguating Syntactic Triggers", *Language Acquisition* 19 (2): 83-143.

Schönenberger, Manuela. 2001. *Embedded V-to-C in Child Grammar: The acquisition of verb placement in Swiss German*. Dordrecht: Kluwer.

Schönenberger, Manuela. 2008. "Three Acquisition Puzzles and the Relation Between Input and Output", in Pedro Guijarro-Fuentes, María Pilar Larrañaga, and John Clibbens (eds.) *First Language Acquisition of Morphology and Syntax*. Amsterdam: John Benjamins, pp. 87-118.

Schwartz, Bonnie and Sten Vikner. 1996. "The Verb Always Leaves IP in V2 Clauses" in Adriana Belletti and Luigi Rizzi (eds.), *Parameters and Functional Heads: Essays in Comparative Syntax*. Oxford: Oxford University Press, pp. 11-62.

Schweikert, Walter. 2005. *The Order of Prepositional Phrases in the Structure of the Clause*. Amsterdam: John Benjamins.

Sells, Peter. 1995. "Korean and Japanese Morphology from a Lexical Perspective", *Linguistic Inquiry* 26 (2): 277-325.

Sells, Peter. 2001. *Structure, Alignment, and Optimality in Swedish*. Stanford: CSLI.

Snyder, William. 2007. *Child Language: The parametric approach*. Oxford: Oxford University Press.

Stowell, Timothy. 1981. *Origins of Phrase Structure*. Ph.D. dissertation, Massachusetts Institute of Technology.

Stromswold, Karin. 1990. *Learnability and the Acquisition of Auxiliaries*. PhD dissertation, Massachusetts Institute of Technology.

Takano. Yuji. 2002. "Surprising Constituents", *Journal of East Asian Linguistics* 11 (3): 243-301.

Tenny, Carol L. 2000. "Core Events and Adverbial Modification" in Carol L. Tenny and James Pustejovsky (eds.) *Events as Grammatical Objects*. Stanford: CSLI, pp. 285-334.

Travis, Lisa deMena. 1984. *Parameters and Effects of Word Order Variation*. Ph.D. dissertation, Massachusetts Institute of Technology.

Vecchiato, Antonella. 2011. *Events in the Grammar of Direct and Indirect Causation*. PhD dissertation, The University of Southern California.

Waldmann, Christian. 2011. "Moving in Small Steps Towards Verb Second: A case study", *Nordic Journal of Linguistics* 34 (3): 331-359.

Waldmann, Christian. 2014. "The Acquisition of Neg-V and V-Neg Order in Embedded Clauses in Swedish: A microparametric approach" *Language Acquisition* 21 (1): 45-71.

Westergaard. 2006. "Triggering V2: The amount of input needed for parameter setting in a Split-CP model", in Adriana Belletti, Elisa Bennati, Cristiano Chesi, Elisa DiDomenico, and Ida Ferrari (eds.), *Language Acquisition and Development: Proceedings of GALA 2005*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 564-577.

Wexler, Kenneth. 1998. "Very Early Parameter Setting and the Unique Checking Constraint: A New Explanation of the Optional Infinitive Stage" *Lingua* 106: 23-79.

Wexler, Kenneth. 2011. “Grammatical Computation in the Optional Infinitive Stage” in Jill de Villiers and Thomas Roeper (eds.), *Handbook of Generative Approaches to Language Acquisition*. New York: Springer, pp. 53-118.

Wexler, Kenneth and M. Rita Manzini. 1987. “Parameters and Learnability in Binding Theory”, in Thomas Roeper and Edwin Williams (eds.), *Parameter Setting*. Dordrecht: Reidel, pp. 41-76.

Wilson, Colin. 2006. “Learning Phonology with Substantive Bias: An experimental and computational study of velar palatalization”, *Cognitive Science* 30 (5): 945-982.

Wurmbrand, Susi. 2005. “Verb Clusters, Verb Raising and Restructuring”, in Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, Volume 5. Oxford: Blackwell, pp. 227-341.

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

Yoon, James Hye Suk. 1994. “Korean Verbal Inflection and Checking Theory”, in Heidi Harley and Colin Phillips (eds.), *MIT Working Papers in Linguistics, Volume 22: The Morphology-Syntax Connection*. Cambridge, MA: Department of Linguistics and Philosophy, MIT, pp. 251-270.

Zwart, Jan-Wouter. 1997. *The Morphosyntax of Verb Movement: A Minimalist Approach to Dutch syntax*. Dordrecht: Kluwer.