DIPLOMARBEIT

Titel der Diplomarbeit

"The Triggering Learning Algorithm and the Problem of Local Maxima"

Verfasser

Andreas Baumann

angestrebter akademischer Grad

Magister der Philosophie (Mag. phil.)

Wien, 2009

Studienkennzahl lt. Studienblatt:	A 328
Studienrichtung lt. Studienblatt:	Allgemeine und Angewandte Sprachwissenschaft
Betreuer:	Ass. Prof. Dr. Martin Prinzhorn

can't we break the cycle in which we reside?
trigger the silence and make up your mind
Eric Kalsbeek

Acknowledgments

In fact, all the people I want to thank are triggers. It is not that they triggered me out of some linguistic structure, but they made it possible for me to start and finish the writing process of this thesis:

Thanks to my supervisor Martin Prinzhorn, especially for giving me free choice of what and how to work on.

Thanks to my colleagues at the Institute of Linguistics and at the Institute of Mathematics for inspiring discussions.

Thanks to my friends, in particular to Jakob and Jo, for non-scientific support in its best sense.

Thanks to my brothers for supporting me in making up this thesis: to Bernhard for assistance with formatting, and to Stefan for proof-reading.

Thanks to my whole family for encouragement of any kind. Particularly, I want to thank my parents, Eva and Peter Baumann, for enabling me to be engaged with the study of linguistics—from first language acquisition to university—and for all the financial and caring support.

Thanks to Theresa for patience, fortitude, and all the rest.

Contents

1 Preface

"Language learnability has been investigated." are the introductory words in [Gold 1967, 447], an influencial information theoretic paper on language learnability. Indeed, theories for natural language acquisition are manifold, but all of them have to account for the same facts: Languages are acquired entirely in merely a few years, every child can learn every language independently of population genetics, children neither get explicit advices about the structure of a grammar, nor are they corrected permanently in case of unacceptable utterances. Moreover, a theory for language acquisition has to explain how such an apparently complex system can be learned, and to factor in why languages are aquired in exactly the way they are learned actually. To find a theory that explicates all these given facts is a main purpose of cognitive linguistics.

In a sense, children learning a language are in a similar position. They are exposed to given facts, that is, sentences, which are uttered by other humans and thus are supposed to be acceptable, and based on their experience they have to find an appropriate theory that can account for the fact that exactly these sentences are well-formed, namely, a suitable grammar.[1]

So, let Gold's introduction be continued:

> *"A class of possible languages is specified, together with a method of presenting information to the learner about an unknown language, which is to be chosen from the class. The question is now asked, 'Is the information sufficient to determine which of the possible languages is the unknown language?'"* [Gold 1967, 447]

Gold's language learning model, and the proof of his theorem in particular, initiated the search for a genuine formal description of language learning processes. In a nutshell, the theorem states that the class of languages, which contains at least one infinite language, cannot be learned using a straight forward learning algorithm as the one sketched in the citation. Since natural language is assumed to be infinite, at first glance Gold's theorem then would entail that the class of natural languages could not be learned by a human learner, which indeed does not represent the human learnability skills.

Needless to say that Gold's—information theoretic—theorem had deep impact on linguisticstraight forwardly literature: It was seen either as formal equivalent to the *poverty of stimulus* argument, hence affirming the existence of innate linguistic knowledge, or as support for *negative evidence* in natural language acquisition. Both implications, however, are on a sticky wicket, as [Gordon 1990] pointed out.

Nevertheless, the desire for an algorithm for language acquisition led to several models, all of which one is the *triggering learning algorithm* (TLA), proposed by Gibson and Wexler (henceforth, GW) in [GW 1994]. This algorithm is a model for the acquisition of grammars in frameworks that rely on *parameters and principles*, such as Government and Binding (GB), HPSG, or LAG. In the TLA the learner moves from one hypothetical grammar to the other dependent

[1] This analogon does not arise by accident: Gopnik and Meltzoff elaborated this idea in [GM 1997] and worked out the so-called *theory-theory*, based on the assumption that children develop theories in cognitive development. See figure 1.

The corresponding formal learning theory, namely *learnability theory* or *inductive inference* in general, is—not surprisingly—based on Gold's investigations.

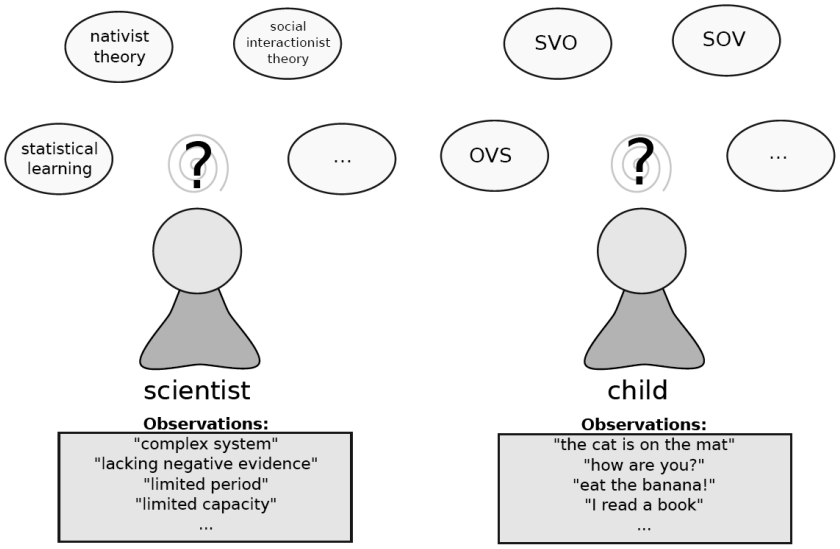


Figure 1: *Schematic representation of theory and language identification, respectively. In fact, it shows that both, theory identification as well as language identification, are two applications of inductive inference.*

on the positive examples, that is grammatical sentences, presented to him—just as in Gold's model, but psycholinguistically motivated.

Niyogi and Berwick (henceforth, NB) almost immediately found a mathematical modelling for the TLA, namely *Markov chains*, presented in [NB 1994]. Using Markov chain theory NB could analyze the algorithm qualitatively as well as quantitativley.

Though there are psycholinguistic motivations for the TLA, there are several problems when applying the TLA to even small sets of parameters, which have been investigated in linguistic research. GW found out that in TLA the learner is likely to get stuck in certain grammars, where he cannot escape, no matter what examples are presented to him. GW refer to such states as *local maxima*. Unfortunately local maxima appear quite often, in particular for grammars which usually can be learned.

The aim of this thesis is to examine the TLA and, in particular, the problem of local maxima. More precisely, I will investigate the solutions for the local maxima problem given so far in literature, and show that most of them are questionable. Then, the approach of including negative evidence into the algorithm will be elaborated. In particular, I will reveal that, if negative evidence is accepted to be part of the language acquisition mechanisms, the problem is solved.

The whole thesis is structured in three parts, each of which is divided into two chapters. Four sections form each of the chapters, which constructively

develop the argumentation.

The first part—the *introduction*—presents the basement for the algorithm, that is, the underlying formal considerations and the grammatical framework within which the TLA operates. The purpose of chapter 2.1 is threefold: First, it constitutes a brief outline of the motivations for learning algorithms in general, and thus also for the TLA. Second, it clarifies questions about the effect of negative evidence, which clears the way for considerations about its implementation, following later on in this thesis. Third, it motivates the assumption of innate linguistic knowledge in a natural way. The innate cognitive structure, *universal grammar*, then is worked on in 2.2. In this chapter, I introduce the grammatical framework, which is focussed on in this thesis, namely a fragment of GB.

The second part—*triggering learning algorithm*—provides a formulation of the TLA, its mathematical modeling, and investigations of its behavior within the introduced grammatical framework. Chapter 3.1 contains the formulation of the algorithm as well as motivations and definitions for all relevant concepts. Moreover, I give a brief introduction into the mathematics required to model the TLA. In chapter 3.1.3 I make use of the modeling in order to analyze the learning behavior of the TLA, which directly leads to the problem of local maxima.

Finally, the third part copes with different *solutions for the local maxima problem*. In 4.1 I utilize NB's mathematical approach to model possible solutions given in literature in order to show their inappropriateness. At last, I will consider negative evidence as a possible solution. Taking the results in 2.1 into account, I develop a formulation for a modification of the TLA, containing negative evidence. Mathematical analysis will show that the modification yields suitable results.

Mathematics used in this thesis mainly comes from theoretical informatics, probability theory, and linear algebra. However, the concepts are fundamental in most cases, and the mathematical toolbox is worked out as we go along, such that no special mathematical knowledge is preconditioned, besides some basic formal principles. Most of the terms can be looked up in the glossary (see appendix B).

The most important informal working hypotheses can be summarized as follows: In general, I assume innate linguistic knowledge in terms of GB theory, as outlined in 2.2. Furthermore, in order to formulate the TLA, I assume the existence of interpretable triggers as in 3.1. Finally, I assume that there exists negative evidence in child directed speech to some extent that can be fully interpreted by the child, which is necessary for 4.2.

One more remark: The TLA is a *model* of language acquisition rather than a *theory*, meaning that the cruical criterion for the quality of the approach is adequacy, not consitency. In particular, what we expect from the TLA being an adequate formal model is that its predictions coincide with empirical data, rather than describing how some cognitive structure is constituted. However, there may be drawn conclusions from its predictions about the nature of this structure. The way the TLA tries to reach adequacy, nevertheless, is to incorporate concepts drawn directly from existing theories of language acquisition.

Following Niyogi [Niyogi 2006, ch. 1.7], a psycholinguistically plausible modeling of natural language acquisition is of interest to several fields of research like psycholinguistics, syntactic theory, computational science, or artificial intelligence. Moreover, concise language acquisition modeling is a precondition

for the computational theory of language evolution, which in turn provides an insight into diachronic syntax and evolutionary theory.

2 Introduction

2.1 Learnability theory

Learnability theory is the mathematical theory about learning formal languages. In general, this area subsumes any kinds of algorithms that enable a machine to learn a language. However, the model of learning handled in this chapter is a so-called inductive inference model, that is, the learner uses information he received so far to make predictions about future events. Originally this model has been developed by the information scientist E. M. Gold, but, what it makes interesting for the topic of natural language acquisition, is the fact that this model is motivated by computational linguistics.

It is clear that a mathematical model of language learning is interesting for computational linguistics but it is less obvious how such a theory can be of use for the theory of natural language learning. Indeed, there has been confusion about Gold's results, which are—admittedly—striking: According to his main theorem, the learnable class of languages does not even contain the least complex language classes in the Chomsky hierarchy.

But, one has to be very precise and cautious to adapt Gold's theory to the theory of natural languages. The main goal of this chapter is to prevent misinterpretations by systematically introducing the learnability model (2.1.1) and in particular an exact formulation of learnability. I will then define the classes of languages which are within the scope of Gold's theorem and which are interesting for linguistic purposes (2.1.2). Thereafter, I will formulate Gold's theorem and show how its proof works in principle (2.1.3), without going into technical detail too thoroughly. Finally, I will work out what relevant conclusions can be drawn from Gold's results, as well as—and this is even more important—which cannot (2.1.4).

One main point of Gold's theorem is the relation between learnability and the kind of information presented to the learner, that is, positive or negative evidence. This point has often been brushed aside in the linguistic literature, mainly to justify innate linguistic knowledge. As we will see, there is no reason to do so. This problem will finally lead us straightforwardly to the second chapter of the introduction part, 2.2. The concept of negative evidence, however, will be of significance for the latter part of this thesis.

2.1.1 Inductive inference

The learnability model formulated by Gold treats the search for a certain target grammar—or language—in a set of grammars. Therefore, it is originally named *language identification in the limit* (cf. [Gold 1967]). More exactly: In the Gold sense a *learnability model* is defined as a triple consisting of a *definition of learnability*, a *method of information presentation* and a *naming relation* [Gold 1967, 449].

Before this triple is examined further, let us introduce some notation: Let $\mathcal{G}$ denote the set of possible grammars, and for each grammar G let $\mathcal{L}(G)$ be the language generated by the grammar (see section 2.1.2 for further details). Furthermore, let G_T denote the target grammar, that is, the grammar, which shall be acquired, and $L_T := \mathcal{L}(G_T)$.

The learning model describes the following situation: At each time t a string

$s_t \in \Sigma^*$ is presented to the learner, where Σ^* denotes the set of possible strings. Note that this string is not necessarily an element of L_T. Whether this is the case, solely depends on the information presentation. The *learner* is defined as a function $\mathcal{F}$, which maps sequences of strings to grammars. Thus, after a time t, the function has value $\mathcal{F}(s_1, ..., s_t) = G_t$.

Language learnability, the first part of the triple, now is defined in the following manner: A target grammar G_T is said to be *identified in the limit*, if after a finite time t_0 all guesses, i.e. values of $\mathcal{F}$, are the same. That is, for all $t \geq t_0$, $\mathcal{F}(s_1, ..., s_t) = \mathcal{F}(s_1, ..., s_{t_0}) = G_T$ – the sequence of hypothesized grammars converges to the target grammar. This means that the learner finally identified the correct grammar corresponding to the target grammar and will never switch to another grammar. A target language L_T is *identified in the limit*, if this is the case for G_T.

Moreover, another definition is made up in [Gold 1967]: A *class of languages* is called identifiable in the limit, if there exists a learner, which is able to identify every language in this class given any sequence of information. It is worth reminding this definition, since it is of great importance for the consequences of Gold's theorem. These consequences will be handled in section 2.1.4.

The second part of the triple is the *method of information presentation*. Gold originally proposed six different methods, which can be devided into the two classes *text* and *informant*. I will not examine the exact definitions of all six presentation methods, as it turns out that in each case three of them are equivalent with respect to the learnability of classes of grammars. Only the basic differences shall be covered here:

1. A *text* is a (probably infinite) sequence of strings in L_T, such that every element of the target language occurs at least once in the sequence.

2. An *informant* is a (probably infinite) sequence of *labeled* stings in Σ^*, such that every element of Σ^* occurs at least once in the sequence.

Here labeled means that the learner receives the information, which of the strings belong to the target language and which do not.[2] In linguistic terms *text* and *informant* can be related[3] to positive evidence and negative evidence together with positive evidence, respectively: In the first case the information presented to the learner consists only of sentences that belong to the target language. In the second case the learner comes to know, which of the sentences are grammatical in the language and which are not.

The third part of the triple finally is the *naming relation*. It assigns names of grammars to languages and Gold distinguishs between two different kinds of relations: *tester* and *generator*. The difference between these two relations is not crucial for the learnability results studied here, for which reason I will not go into details.

Following Gordon [Gordon 1990], two remarks have to be made on Gold's model:

[2] Formally this can be described by a mapping $l : \Sigma^* \to \{0, 1\}$, where l is the characteristic function of L_T. The informant now is a sequence of ordered pairs $(s, l(s))$ for $s \in \Sigma^*$.

[3] It is important to note that information presentation by informant is *not equivalent* to what is expected to be negative evidence in linguistics. Information presentation by informant is much stronger than corrective feedback. I will return to this difference later in this thesis.

1. *"Given the characteristics of the learner just described, it should be clear that* learning *within this paradigm is not what we normally associate with this term.* [...] *The phrase, "in the limit" here denotes the criterion for success."* [Gordon 1990, 218]. Gordon points out that it might be possible, that already the first guess leads to the correct target grammar. On the other hand, also a large number of examples could be presented before the right guess. Within this model the number of required examples is not bounded (upwards) and is of no significance for the learnability of a language. It is obvious that this definition of learnability differs from what would be expected in a natural language learning setting.

2. *"In addition, note that for learnability* [of a class of languages] *to be guaranteed in every case, the learner must hear* all *of the sentences in the language for text presentation, plus* all *of the sentences that are not in the language (appropriatly labeled) for informant presentation."* [Gordon 1990, 218]. This is, because for any language there exists another (more extensive) language within the same class, which requires more information to be learned. To learn the *whole class*, then, the number of required sentences must be infinite.

Now, before Gold's theorem on the learnability of languages is stated, I will recapitulate the different classes of (formal) languages in the following section.

2.1.2 Chomsky hierarchy

Formal grammars are ordered in the so called Chomsky hierarchy, and there exists an extensive debate in literature about where grammars of natural languages should be arranged in this ordering (cf. [Chomsky 1957], [Kracht 2003, 2.7]). The classes in the hierarchy and the answer to the previous question are important for the result of Gold's considerations.

First of all, let us briefly review the notion of a formal grammar: A *formal grammar* is a quadruple $G = \langle S, N, \Sigma, R \rangle$, where S is the *start symbol*, N is the *set of nonterminal symbols* and Σ is the *terminal alphabet.* The *Kleene closure* Σ^* for a given alphabet is defined as the set of all possible strings that can be formed by concatenation using the elements of Σ. Furthermore, R is a finite *set of rules* of the form $a_1 n a_2 \to a_3$, where $n \in N$ and $a_i \in (\Sigma \cup N)^*$. This means that a string consisting of at least one nonterminal symbol will be derived into another string, which may or may not contain nonterminal symbols.

The *formal language generated by* G, $\mathcal{L}(G)$, then is the set of all strings $\sigma \in \Sigma^*$ that can be derived from S by the given rules and the elements of N and Σ. Thus, for a given alphabet Σ, a *formal language* generally can be defined as a certain subset of Σ^*. It is worth to remark that two different grammars can yield exactly the same formal language, meaning that if there is a rather complex grammar for a given language, it may well be the case that one can find a simpler one that generates the same language. The Chomsky hierarchy states what terms such as "complex" or "simple" mean for a formal grammar (cf. [Hausser 2000, 160], [Kracht 2003, 54]):

1. The most complex grammars are *unrestricted* or *type 0* grammars. Any formal grammar is of this type: This means that the rules of any grammar are of the form $a_1 n a_2 \to a_3$. Unrestricted grammars are the most complex

ones, since there is no restriction on the rules unlike in all other types of grammars. These grammars generate *recursively enumerable* languages, that is, languages that can be calculated by a *Turing machine*. It suffices to notice that such grammars are far too unrestrictive to describe natural languages.

2. *Context sensitive* or *type 1* grammars have rules of the form $a_1 n a_2 \to a_1 \gamma a_2$, where $n \in N$, $a_1, a_2, \gamma \in (\Sigma \cup N)^*$. That is, the context of a certain nonterminal symbol remains the same but the part in between can be substituted by any other string of nonterminal or terminal symbols ($\gamma = \varepsilon$, where ε denotes the empty string, is not allowed). Languages generated by a context sensitive grammar can be calculated in *exponential* time. It is important to note that derivations never get shorter.

3. *Context free* or *type 2* grammars are context sensitive and all rules are of the form $n \to \gamma$, where $\gamma \in (\Sigma \cup N)^*$. Such grammars are more restricted since only one nonterminal symbol is allowed on the left hand side: Unlike context sensitive rules, the derivational rules in type 2 grammars have a limited field of vision and cannot see beyond their scope. It takes *polynomial* time to state if a string belongs to a language generated by a context free grammar.

4. The most restricted grammars in the Chomsky hierarchy are *regular* or of *type 3*. A grammar is said to be regular if it is context free an all rules are of the form $n \to a$, where $n \in N$ and $a \in (\Sigma \cup \{\varepsilon\}) \times (N \cup \{\varepsilon\})$. Thus, the right hand side of a rule only consists of at most two symbols: a terminal followed by a nonterminal, a single terminal or a single nonterminal (or ε). Regular languages can be calculated in *linear* time by a *finite state automaton*. However, grammars of natural languages are more complex than regular grammars: *"Any attempt to construct a finite state grammar for English runs into serious difficulties and complications at the very outset* [...]. *That is, it is* impossible *not just difficult, to construct a device of the type described above* [...] *which will produce all or only the grammatical sentences of English."* [Chomsky 1957, 20-21].

In general we call a formal language *recursively enumerable*, *context sensitive*, *context free* or *regular* if it is generated by a grammar of type 0, 1, 2 or 3, respectively. All recursively enumerable languages form the *recursively enumerable class* of languages, the context sensitive languages form the *context sensitive class* and so on. It is clear from the definitions that these classes form proper subsets of the class of lower type in each case. Thus, the regular class is a subset of the context free class, which in turn is a subset of the context sensitive class and so forth.

Moreover, it is easy to see that the complexity of a language only depends on how the rules are stated in the generating grammar. The cardinality of Σ, N or R is not relevant in this formal context, but it may well be in a natural language learning situation (meaning the number of words in the lexicon or of different rules).

Two more classes of languages shall be mentioned here, as they are important for Gold's theorem:

Table 1: Classes of languages

Class	*Languages*	*Type*	*Grammar*	*Complexity*
recursively enumerable	rec. en.	type 0	unrestricted	undecidable
context sensitive	cont. sens.	type 1	cont. sens.	exponential
context free	cont. free	type 2	cont. free	polynomial
regular	regular	type 3	regular	linear
superfinite	all fin. + inf.	-	regular	linear
finite cardinality	all fin. lang.	-	listing	linear

1. A class of languages is called *superfinite* if it contains "all languages of finite cardinality and at least one of infinite cardinality." [Gold 1967, 452].

2. At last there is the class of *finite cardinality languages*, consisting of all languages of finite cardinality. Such a language is nothing but a finite listing of all grammatical sentences, e.g., the Boston telephone directory (cf. [Gordon 1990], [Pinker 1981]).

For convenience, all classes of languages shall be summarized in table 1.

Now, it is widely assumed, as argued above, that the class of natural languages is more complex than regular grammars and less complex than unrestricted grammars. Where exactly natural languages have to be ranked in the Chomsky hierarchy still remains an open question (see [Lasnik 1981] or [Kracht 2003, 2.7] and citations therein), however, this problem shall not bother us working on learnability results. Two facts about the class of natural languages are crucial for Gold's theorem: First, this class consists of at least one infinite language. Hence, we could speculate that natural language class is at least superfinite. Second, natural languages are more restricted than recursively enumerable ones.

2.1.3 Gold's theorem

I am now ready to state Gold's theorem. More exactly it is not only one but several different propositions of which I will cover only those, which are important for natural language learnability. All others provide technical details only of interest for theoretical informatics and mathematical logic. Besides, I will simplify the theorems for the following reason: All kinds of information presentation by informant turn out to be equivalent [Gold 1967, thm I.3] as well as all kinds of information presentation by text for our purposes ([Johnson 2004], see comment later on). Thus, I will not distinguish between different types in each case. Furthermore, I will make no differences between the naming relations tester and generator (see [Gold 1967, thm I.1], [Johnson 2004, (A6), (A7)]).

The following facts are modified formulations of the results (A5)-(A7) in [Johnson 2004] in order to stick together with table 1:[4]

[4]This simplification can be justified in the following way: The class of *recursive* languages (mentioned by Gold) is a subset of the class of recursively enumerable languages. Thus, if the latter is not identifiable in the limit, then this also holds for the former. The argumentation

Table 2: *Learnability results*

Class	*Learnable by*
recursively enumerable	-
context sensitive	informant
context free	informant
regular	informant
superfinite	informant
finite cardinality	text, informant

(1) *Informant based learning:* Using information presentation by informant the class of context sensitive languages is learnable but the class of recursively enumerable languages is not.

(2) *Text based learning:* Using information presentation by text the finite cardinality class is learnable but the superfinite class is not.[5]

The results are summarized in table 2.

Fact (2) is what is referred to as Gold's theorem and the idea of its proof is the following (see [Johnson 2004] for an extensive explanation): Imagine an infinite sequence of finite languages $L_1, L_2, L_3, ...$ with the following property:

(3) $L_1 \subseteq L_2 \subseteq L_3 \subseteq ....$

Furthermore, let L_∞ be the language that consists of exactly those elements which are in $L_1, L_2, L_3, ...$ Let $L_\infty, L_1, L_2, L_3, ...$ form the class C. This clearly is a subclass of the superfinite class, since all languages but L_∞ are finite.

Assume that C is learnable, that is, all languages in this class are so, too. Now, positive examples from the target language, say L_1, are presented to the learner such that after some time he will identify the target language. If examples from L_2 are added to the information sequence, the learner will identify L_2 and so on. If the learner wants to identify L_∞, he runs into serious problems, since all elements of this language are also contained in one of the L_i. Hence, this language cannot be identified, a contradiction.

It is important to emphasize that fact (2) makes assertions about the learnability of *classes* not of single languages. [Johnson 2004] reveals several misinterpretations of Gold's theorem in the literature and this property of Gold's statement (namely the restriction to classes) has been overlooked multiple times (see for instance [PS 2006, ch. 3]). As Johnson states, *"any class containing exactly one language would be trivially learnable, because there exists a constant function that always guesses the language"* [Johnson 2004, 578]. Or, as Gordon put it,

---

is analogous for *primitive recursive* languages, which are a superset of the context sensitive class.

[5] To be precisely, one has to exclude the combination of primitive recursive text and the generator naming relation from this proposition. Johnson made up the following formulation: *"(A7)* An anomaly: *Using Primitive Recursive text and generator naming relation, the class of Recursively Enumerable languages is identifiable in the limit"* ([Johnson 2004, 590]) and hence are all other classes in table 1. Anyway, he also states that *"there doesn't appear to be any interesting psychological interpretation of this fact."* ([Johnson 2004, 583])

> *"*[w]*hether a language is in a learnable or an unlearnable class, then, says nothing about whether the language itself is potentially learnable by humans. What Gold's (1967) proof demonstrates is whether each language in a class is destinguishable from other languages in the class, on the basis of input."* [Gordon 1990, 218]

One last remark: According to the definition of learnability given in 2.1.1 a class of languages is learnable if there exists a learner such that for any given information sequence and for any given language of the class, the learner is able to identify this language. Thus, if a class is unlearnable this means that for all learners the following holds: There is a sequence and there is a language, which cannot be learned using this sequence. Or equivalently: For every learner we can find a certain linguistic environment where he cannot learn a given language of the class. Johnson states that *"child language acquisition may be restricted by Gold's Theorem, but this restriction only applies to cases that don't occur in any normal environment, and thus have no practical significance"* [Johnson 2004, 587].

The purpose of the remark above is to show that one has to be careful when applying mathematical models to—here: linguistic—problems. What implications Gold's theorem might evoke will be the topic of the following section.

2.1.4 Implications for linguistics

Through the preceding sections we have seen how language learning can be modeled straightforwardly (*language identification in the limit*) and how the learnability of classes of languages is restricted. This—namely the fact that only the finite cardinality class is learnable without labeled information—was originally shown by Gold and summarized in table 2.

Since children actually do learn natural languages, Gold's theorem (see facts (1) and (2) in 2.1.3) suggests one or more of the following logical consequences:

(4) The class of possible natural languages is smaller than the context sensitive (or context free, or regular) class,

(5) Gold's model of language learning is not appropriate to explain natural language learning,

or finally

(6) there is some kind of negative evidence in language acquisition such that information presentation by informant can be guaranteed.

To understand the first consequence it is useful to quote Gold's explanation:

> *"The class of possible natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context sensitive, it is not true that any context sensitive language can occur naturally.* [...] *In particular, the results on learnability from text imply the following: The class of possible natural languages, if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality."* [Gold 1967, 453]

Indeed, there are no natural languages of finite cardinality. As Gordon (and originally Pinker, see [Pinker 1981]) pointed out, an example for a finite language would be the Boston telephone directory, *"which only the most accomplished mnemoist could actually learn"* [Gordon 1990, 218], although it belongs to the finite cardinality class learnable by text.

But, only because natural languages are of higher *complexity* than finite cardinality languages and because of the context sensitive (or context free, or regular) class being a superset of the finite cardinality class, this does not automatically mean that the class of natural languages must contain all languages of lower complexity, in particular it is not of logical necessity that it contains all finite languages. Lasnik put it this way: *"Another consequence* [of the incorrect conclusion] *is the failure to recognize that there must be restrictions 'from below' as well as 'from above'"* [Lasnik 1981, 4].

This, in turn, is also due to the presented modeling of language acquisition, which is the point of the second consequence: It may well be that Gold's theorem and furthermore the whole mathematical modeling cannot be adopted to the natural language learning problem. For the first point, i.e. the theorem itself, one might critisize that the property of containing all finite cardinality languages is beyond the scope of any linguistically reasonable class of natural languages, thus, Gold's theorem being worth nothing in this context.

For the second point, it is allowable to claim that the definition of learnability made up by Gold does not represent natural language learning satisfactorily. In particular, it is not realistic that a learner acquires a language luckily by the first guess (see end of section 2.1.1). Note that Gold's theorem only makes statements about whether a language is learnable or not, not about how long it takes.

In order to avoid confusion, Johnson differentiates between the following two definitions of learnability (cf. [Johnson 2004, 585]):

(7) *Identifiability*: A class C of languages is identifiable if there is a learner that can identify each language L in C in the limit (in Gold's sense), given any information sequence.

(8) *Acquirability*: A class C of languages is acquirable if every normal human learner can learn approximately every language L in C within a certain period of time, given any information sequence.

Clearly, definition (8) is informal but it comprises important changes: The quantifier here ranges over all learners, that is, only because there exists one mnemoist, which is able to learn any telephone directory, this does not mean these are learnable languages. Contrary, if someone is not able to learn a natural language because of an inherited genetic disorder, we would not say that this language is not learnable in general. In the definition *approximately* means that it suffices for the learner to acquire a language which is close to the target language. Furthermore, language acquisition now is restricted to *a certain period of time*, say 12 years. Note that both definitions are logically independend of each other.

The third possible consequence of Gold's theorem is the existence of some type of negative evidence, which is equivalent to information presentation by informant. If it exists, then all classes of languages which come into consideration when classifying natural languages regarding their complexity are learnable.

Gold suggests that *"[t]he child may receive the equivalent of negative instances for the purpose of grammar acquisition when it does not get the desired response to an utterance"* [Gold 1967, 454].

It is widely assumed that children acquire languages with merely no negative data, such as corrective feedback (see for instance [Hyams 1986], [Pinker 1981] or [Lasnik 1981]). However, there also exist considerations about alternative ways to provide negative feedback to children (cf. [BS 1988]). The topic of negative evidence will remain of interest for this thesis, but for the time being I will omit this possibility. It will be worked on more extensively in section 4.2.

Due to the missing negative feedback another consequence of Gold's theorem is mentioned in literature, namely that there are innate constraints on the language acquisition process. Though this circumstances are referred to as *the logical problem of language acquisition* (e.g. [Grimshaw 1981], from now on LPLA), the assumption of innate knowledge is *not* a logical implication of Gold's theorem like the three consequences above. [Johnson 2004, 583] structured the line of argumentation in the following way:

1. *"If there are no constraints on language acquisition, then either children have access to negative evidence or natural languages are unlearnable.*
2. *If they exist, the constraints in question must be innate.*
3. *Children don't have access to negative data.*
4. *Natural languages are learnable.*
5. *∴There are innate constraints on language acquisition."*

Although this reasoning is correct as well as meaningful[6], it is logically independend of Gold's statements.

The scope of this chapter was to introduce all concepts necessary to fully understand the theorem of Gold, containing a definition of Gold's learning model as well as a description of different classes of grammars. Then, I gave a formulation of the theorem and an outline of its proof. At last, I discussed what relevant implications can be made for linguistic research and how Gold's theorem is related to LPLA .

How this problem is accounted for in linguistics—or more precisely: in syntactic theory—is the subject of the next chapter. But before I close this rather theoretical and mainly mathematical part about inductive inference, I would like to cite the somehow warning words of Gordon [Gordon 1990, 219]:

> *"[T]he issue of whether a class of languages is learnable from text or informant presentation is quite orthogonal to the issue of whether learning those languages requires innate knowledge. Proving that a particular class of languages is identifiable by a Turing machine does not tell us whether a quite general cognitive structure of the kind that might exist in humans could induce the correct grammar."*

[6] Note that in this deduction point 1 is a not proven assertion, which is similar—though not equivalent—to Gold's theorem. In particular, it is not a *logical* necessity that if there are no constraints, then there is negative evidence or natural languages are learnable. Rather, it is an empirical relation between these facts.

2.2 Grammatical framework

Through the previous chapter I have shown how Gold's theorem can deliver meaningful implications and consequences for linguistics. Moreover, I have introduced LPLA and its consequence that there must be innate constraints on language acquisition, due to the (nearly complete) absence of negative evidence. That these constraints are more than restrictions on the set of possible grammars *"'from below' as well as 'from above'"* [Lasnik 1981, 4] in terms of classes, as mentioned, is one main focus of the theory of generative grammar, which I will introduce in this chapter.

Although there are theories of phonology or morphology acquisition (see examples in [Niyogi 2006]), I will—just as GW—concentrate on syntactic theory in this thesis. More precisely, I will focus on *phrase structure grammars* (PSG), though there exist acquisition theories working with *lexical functional grammars* (LFG, [Pinker 1981]) or *categorial grammars* (CG, [Briscoe 2002b]). This decision just is a pragmatical one: First, PSG are the most common in syntactic theory [Chomsky 1957], and second, GW used PSG to applicate their algorithm. Therefore, it will be easier to draw comparisons. Nevertheless, any other formal theory of grammar would have served as well.

In the first section I will present an overview about how LPLA affects on linguistic theory, and introduce the concept of *universal grammar*. This concept will be elaborated further in sections 2.2.2 and 2.2.3 by defining several *principles* and *parameters*. Finally, I will close the chapter by introducing a linguistic space of grammars using three parameters. This space will be the foundation for applications of the TLA, which will be formulated and investigated in the second part of this thesis.

2.2.1 Universal grammar

The approaches to LPLA are manifold. Be it the assumption of direct or indirect negative evidence in child directed speech [BS 1988], or be it the supposition that—seen in an evolutionary context—only the learnable parts of a grammar are actually learned [Zuidema 2003]. The main approach in linguistic theory, however, is to assume innate linguistic knowledge.

Whether such innate knowledge is to be equated with the implication of Gold's theorem, that the class of possible grammars must be restricted (see for instance [Nowak 2001]), remains questionable. I discussed this relation in section 2.1.4. Nevertheless, it is not only the absence of regular *direct* negative evidence (for a discussion on several kinds of negative evidence see [BS 1988], [BMT 1995], [Valian 2009], or [Gordon 1990], summarized in this thesis later on), but also other thought-provoking facts about primary linguistic data:

> *"The data base to which the child is exposed is impoverished in two respects. The data are both 'degenerate' and 'deficient'. They are degenerate insofar as they contain performance errors, for example, slips of the tongue, false starts, and so on. From this point of view of acquisition, the deficiency of the data is much a more serious problem. The data are deficient in that the child receives no direct evidence of ambiguity, synonomy, or ungrammatical sentences."* [Hyams 1986, 23f]

Note that these problems are not captured by Gold's acquisition model. The fact that the data presented to the child is faulty in this sense, is also called the *poverty of stimulus*.

This innate linguistic knowledge is in literature referred to as *language acquisition device* (henceforward abbreviated LAD). The LAD contains all, which is necessary to acquire a natural language, and though there are several proposals about the exact form of the LAD (see [Grimshaw 1981] for a comprehensive explanation), the main idea is that a *"model of LAD must incorporate a theory of UG with an associated finite set of finite-valued parameters defining the space of possible grammars, a parser for these grammars, and an algorithm for updating initial parameter settings* [...]. *It must also specify the starting point for acquisition"* [Briscoe 2002a, 3].

UG stands for *Universal Grammar*, that is to say a grammar, which is powerful enough to potentially generate all natural languages. To be specific, UG consists of variable rules, often called *principles*, that is, for every child the potentially possible grammatical rules are the same, but they are restricted and altered by so-called *parameters*, which are set to determine a certain grammar.

The function of UG is twofold: First, it provides an explanation for linguistic diversity, meaning that UG can yield every grammar for a natural language. Second, it allows a child to acquire a language more systematically. Remember that in Gold's setting the space of possible grammars is unordered insofar as there is no specific relation between two grammars. In terms of UG we can differentiate between grammars by their parameter settings.

In the next three sections I will elaborate definitions for several principles and parameters as for linguistic parameter spaces, that is, spaces of grammars.

2.2.2 Principles

As we have seen, in UG there are two parts, *parameters* and *principles*, of which this section treats the latter ones. The theory of syntax itself is separated into several subtheories such as θ-theory, case theory or X-bar theory. For the purpose of modeling language acquisition I will only handle a fragment of UG, namely X-bar theory, which is the basic theory about word ordering. Only few other principles will be mentioned here, mainly to stick together with the data in [GW 1994].

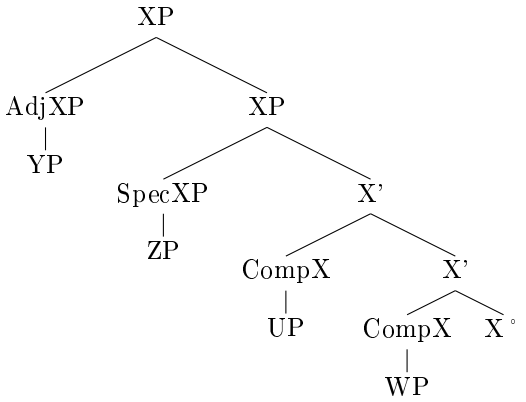
Basically X-bar theory can be written as formal phrase structure grammar, as described in 2.1.2. The rules then would be as given in (9), where X and Y are elements of $\{V, I, C, D, Adv, A, P, N\}$ and where **S** denotes the starting symbol.

(9) *Phrase structur grammar rules for X-bar theory*

1. **S**→XP
2. XP→SpecXP X'
3. XP→AdjXP XP
4. X'→CompX X°
5. AdjXP→YP
6. SpecXP→YP
7. CompX→YP

The basic phrase structure generated by this rules can be represented by the following tree (clearly, it can be enlarged by several adjuncts or complements):

(10)



An implicit principle of the grammar in (9) is that each X° projects to a maximal expansion XP. However, (9) is overgenerating, since there is no restriction on which categories may or may not be inserted. Thus, even sentences without a verb would be accepable. For simplicity, let us assume that the only available categories are C (complementizer), I (inflection) and V (verb), and that the underlying ordering is $[_{CP}...[_{IP}...[_{VP}...]]]$. This could be reached by making up restricting rules such as *S→CP*, *CompC→IP* or *CompI→VP*.

Furthermore, let adjuncts be restricted to CPs, and let there be the following rules:

(11) *Base generating categories*

1. I° →Aux (for *auxiliary*)
2. I° →Ø
3. C° →Ø
4. CompV→O (for *object*; or O1 and O2, if there is a direct and an indirect object)
5. SpecVP→S (for *subject*)
6. V° →V (for *verb*)
7. AdjCP→Adv (for *adverb*)

So, it can be seen where the different categories are base generated within this fragment of syntactic theory. It has to be remarked that these rules should substitute, not just enhance, some of the rules given in (9).

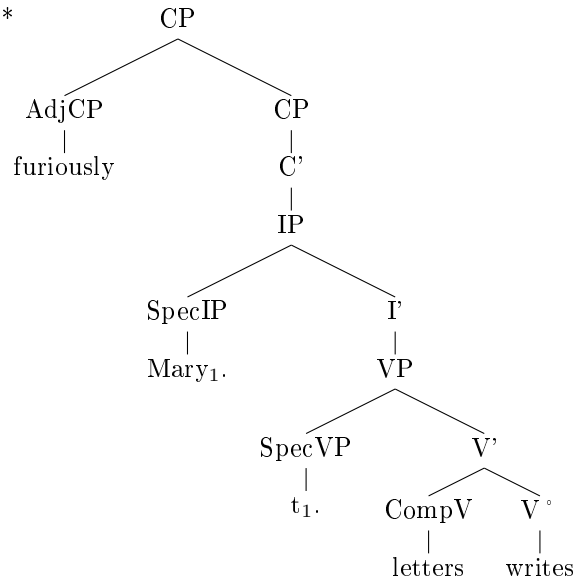
Two additional rules are necessary to explain the possible sentence structures in [GW 1994]: First, every subject has to move obligatory to SpecIP. This is referred to as A-movement and is a matter of case theory, therefore I will not go into detail here.[7] Second, the number of objects depends on the valence of a verb, or more precisely, on the number of its θ-roles. Anyway, this is beyond the

[7] In principle, we could equivalently state that the subject is base generated in SpecIP. Within this restricted framework this would not change anything, though it would change a lot if adjoining more principles. In order to stick to the general approach, I leave it as described above.

scope of X-bar theory and shall be treated as a given here (see [Hyams 1986] for a more elaborated investigation).

The rules described here nearly define a (rather restricted) grammar of English. If applied to English sentences in the just presented form, we would get the uncorrect prediction (t stands for a trace):

(12) *



However, there has only one rule to be changed in (9) to get the right word order *furiously Mary writes letters*. We only have to substitute $V' \rightarrow Comp V\ V^\circ$ by the new rule $V' \rightarrow V^\circ\ Comp V$. This is the point where *parameters* come into play, which will be focussed on in the next section.

2.2.3 Parameters

While principles are assumed to be the same in all natural languages, this does not hold for parameters. More precisely, this means that in all grammars we can find the same parameters as such, but they differ from each other in their settings. Parameters are thought of as binary variables, that is, a parameter can be set in exactly two different ways. Formally, we can define the *parameter value* as a function p which maps each parameter to either 0 or 1.[8] This does not necessarily mean that if a parameter has value 0, it is "set off". In some cases such an interpretation would not be meaningful.

In this section I will cover four parameters—*spec*, *comp*, *V2* and *null subject*—of which the first two are closely related to the principles mentioned in the previous section.

To begin with, the *spec* parameter determines whether the specifier specXP of a phrase or its projection X' comes first. That is, rule 2 in (9) is modified in the following manner:

(13) *Spec parameter*

[8] Thus, for a certain parameter x, we have either $p(x) = 0$ or $p(x) = 1$. For simplicity, I will henceforth write p_x for $p(x)$. If the parameters are ennumerated, x is often just the number assigned to the parameter.

Table 3: *Basic word orders and their parameter settings*

tree in (15)	p_{spec}	p_{comp}	*word order*
1	0 (*spec-first*)	0 (*comp-first*)	SOV
2	1 (*spec-final*)	0 (*comp-first*)	OVS
3	0 (*spec-first*)	1 (*comp-final*)	SVO
4	1 (*spec-final*)	1 (*comp-final*)	VOS

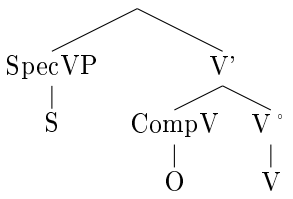
XP→X' SpecXP iff $p_{spec} = 1$ (*spec-final*)
XP→SpecXP X' iff $p_{spec} = 0$ (*spec-first*)

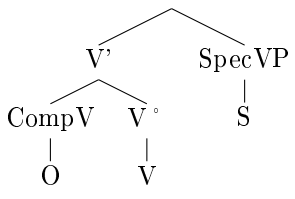
Similarly, the *comp* parameter determines the position of the complement, which is a revision of rule 4 in the list.

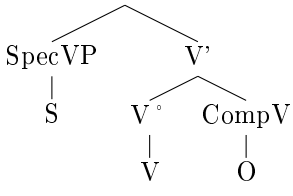
(14) *Comp parameter*
X'→X° CompX iff $p_{comp} = 1$ (*comp-final*)
XP→CompX X° iff $p_{comp} = 0$ (*comp-first*)

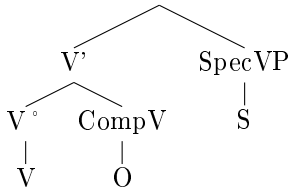
Looking at a VP the following four basic word orders can be derived, which are summarized in table 3.

(15) *Basic word orders*

1.


2.


3.


4.


An example for a SVO language is English, thus we see that if the comp parameter is set to 1, i.e. if the complement follows the verb, we get the right word order in example (12) in the previous section.

It has to be remarked that—to simplify matters—I assume that both parameters only apply to the categories I and V. Thus, within the CP there remains the fixed order as given in (9). Following Gibson and Wexler *"*[t]*his simplification will not affect any of the nonlearnability results to follow, because the learnability problem is at least as difficult (and possibly more difficult) when further parameters are added to a parameter space, when considering the same sentence patterns"* [GW 1994, 422].

A parameter affecting the movement of some elements and thus having an influence on the word order is the *verb second* parameter, henceforth abbreviated as V2. In V2-languages, such as German, in declarative sentences the finite verb moves to C° and then the position SpecCP has to be filled. Since in our restricted framework, there is yet no distinction between finite and infinite verbs, I now assess that in sentences with Aux and V, the auxiliary is finite, and in sentences without Aux, V is finite.

For our purposes this also means that either S, O1 or O2 is moved to SpecCP unless there is an adverbial adjunct in the sentence. We define the value of V2 as follows:

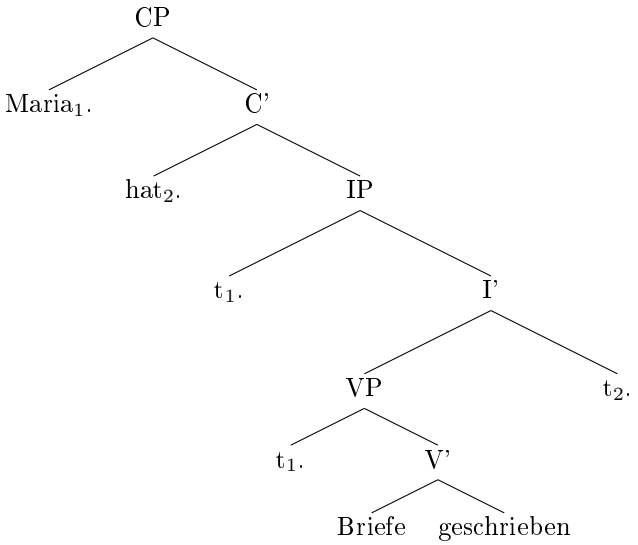
(16) *V2 parameter*

finite verb moves to C° iff $p_{V2} = 1$

finite verb does not move iff $p_{V2} = 0$

For a V2-language like German ($p_{comp} = 0$, $p_{spec} = 0$, $p_{V2} = 1$) we thus obtain sentence structures as in (17).

(17)



Maria	*hat*	*Briefe*	*geschrieben.*
Mary	have [finite]	letters	write [infinite]

'Mary has written letters'

The last parameter to discuss is the *null subject* parameter. Let it be illustrated by the following Italian example :

(18) *"Mangia una mela*

'Eats an apple' " [Hyams 1986, 31]

It is widely assumed that this sentence is equivalent to sentence (19), that is, we expect that there is a not phonetically realized pronominal subject (therefore often also called *pro-drop* parameter).

(19) *"Lui (lei) mangia una mela*
'He (she) eats an apple' " [Hyams 1986, 31]

I will not go into detail about how exactly this omission of pronominal subjects is caused (in fact, in the fragmental theory discussed here we do not even differentiate between pronominal and non-pronominal subjects), but *"there is a wide consensus that the pro drop phenomenon is related to properties of the INFL node"* [Hyams 1986, 32]. From now on, I will determine the values of the null subject as in the definition to follow.

(20) *Null subject parameter*
S→Ø is a rule iff $p_{NS} = 1$
This rule is applied after all movements.

Remark that only because of the existence of an additional rule, there is no necessity to use it. Hence, it is optional, which guarantees, that both Italian sentences (18) and (19) can be generated.

Three of the just specified parameters will be used to span a parameter space, as stated in the section below. The forth parameter, namely the null subject parameter, will remain of interest for the so-called *subset problem*, which I will explain in section 4.2.4.

2.2.4 A linguistic 3-parameter space

Let us now use the parameters *spec*, *comp* and *V2*, to introduce and illustrate the concept of a *linguistic parameter space*. Generally, by this term we understand the set of possible grammars, let it be denoted by $\mathcal{G}^n$, which can be generated by a certain set of parameters $\{p_1, \ldots, p_n\}$, for a given natural n. As we assume, that all linguistic parameters are binary valued, we have a total of 2^n possible grammars in $\mathcal{G}^n$. Since all of the p_i equal either 0 or 1, we can identify each grammar with a finite sequence of ones and zeros.

For three parameters, take $p_1 = p_{spec}$, $p_2 = p_{comp}$ and $p_3 = p_{V2}$ for instance, we thus get 8 grammars, subsumed in

(21) $\mathcal{G}^3 = \{(000), (001), (010), (011), (100), (101), (110), (111)\}$,

where (010) means *spec first*, *comp final*, *not V2*, for example. This just describes the parameter setting for English in our restricted framework.

Finally, to complete the notion of a space, we introduce a distance measure on $\mathcal{G}^3$, the so-called Hamming distance [Bronstein 2005, 625]. For two binary sequences s_1, s_2, we define the Hamming distance $d_{hamm}(s_1, s_2)$ to be the number of positions for which s_1 differs from s_2.

So, it can be seen that (001) is farther away in this sense from (010) as (000) is. We get $d_{hamm}((001), (010)) = 2$ and $d_{hamm}((001), (000)) = 1$. Here, (001) is the parameter setting for German and (000) constitutes a setting for Hindi (see [Niyogi 2006, 125]). This demonstrates, how even languages of the same family can be distant in terms of parameter settings.

Using the phrase structure rules of section 2.2.2, several sentence patterns can be derived for each parameter setting in $\mathcal{G}^3$. Following GW, there are 12 main categories (cf. [GW 1994, 422f]):

(22) *Sentence patterns*

Table 4: List of languages in a 3-parameter space (cf. [GW 1994, 424] and [Niyogi 2006, 125])

Language	p_{spec}	p_{comp}	p_{V2}	*Sentence patterns*
L_1	1	1	0	"V S", "V O S", "V O1 O2 S", "AUX V S", "AUX V O S", "AUX V O1 O2 S", "ADV V S", "ADV V O S", "ADV V O1 O2 S", "ADV AUX V S", "ADV AUX V O S", "ADV AUX V O1 O2 S"
L_2	1	1	1	"S V", "S V O", "O V S", "S V O1 O2", "O1 V O2 S", "O2 V O1 S", "S AUX V", "S AUX V O", "O AUX V S", "S AUX V O1 O2", "O1 AUX V O2 S", "O2 AUX V O1 S", "ADV V S", "ADV V O S","ADV V O1 O2 S", "ADV AUX V S", "ADV AUX V O S", "ADV AUX V O1 O2 S"
L_3	1	0	0	"V S", "O V S", "O2 O1 V S", "V AUX S", "O V AUX S", "O2 O1 V AUX S", "ADV V S", "ADV O V S", "ADV O2 O1 V S", "ADV V AUX S", "ADV O V AUX S", "ADV O2 O1 V AUX S"
L_4	1	0	1	"S V", "O V S", "S V O", "S V O2 O1", "O1 V O2 S", "O2 V O1 S", "S AUX V", "S AUX O V", "O AUX V S", "S AUX O2 O1 V", "O1 AUX O2 V S", "O2 AUX O1 V S", "ADV V S", "ADV V O S", "ADV V O2 O1 S", "ADV AUX V S", "ADV AUX O V S", "ADV AUX O2 O1 V S"
L_5	0	1	0	"S V", "S V O", "S V O1 O2", "S AUX V", "S AUX V O", "S AUX V O1 O2", "ADV S V", "ADV S V O", "ADV S V O1 O2", "ADV S AUX V", "ADV S AUX V O", "ADV S AUX V O1 O2"
L_6	0	1	1	"S V", "S V O", "O V S", "S V O1 O2", "O1 V S O2", "O2 V S O1", "S AUX V", "S AUX V O", "O AUX S V", "S AUX V O1 O2", "O1 AUX S V O2", "O2 AUX S V O1", "ADV V S", "ADV V S O", "ADV V S O1 O2", "ADV AUX S V", "ADV AUX S V O", "ADV AUX S V O1 O2"
L_7	0	0	0	"S V", "S O V", "S O2 O1 V", "S V AUX", "S O2 O1 V AUX", "ADV S V", "ADV S O V", "ADV S O2 O1 V", "ADV S V AUX", "ADV S O V AUX", "ADV S O2 O1 V AUX"
L_8	0	0	1	"S V", "S V O", "O V S", "S V O2 O1", "O1 V S O2", "O2 V S O1", "S AUX V", "S AUX O V", "O AUX S V", "O1 AUX S O2 V", "O2 AUX S O1 V", "ADV V S", "ADV V S O", "ADV V S O2 O1", "ADV AUX S V", "ADV AUX S O V", "S AUX O2 O1 V", "S O V AUX", "ADV AUX S O2 O1 V"

1. Some order of V and S.
2. Some order of V, S and O: two patterns if $p_{V2} = 1$, one if $p_{V2} = 0$.
3. Some order of V, S, O1 and O2: three patterns if $p_{V2} = 1$, one if $p_{V2} = 0$.
4. Some order of V, Aux and S.
5. Some order of V, Aux, S and O: two patterns if $p_{V2} = 1$, one if $p_{V2} = 0$.
6. Some order of V, Aux, S, O1 and O2: three patterns if $p_{V2} = 1$, one if $p_{V2} = 0$.
7. Adv followed by some order of V and S.
8. Adv followed by some order of V, S and O.
9. Adv followed by some order of V, S, O1 and O2.
10. Adv followed by some order of V, Aux and S.
11. Adv followed by some order of V, Aux, S and O.
12. Adv followed by some order of V, Aux, S, O1 and O2.

The sentence patterns here describe *degree-0* sentences, that is, unembedded declarative sentences. For an acquisition theory of higher degree sentences see [Wexler 1981] and [Williams 1981]. All possible patterns shall be summarized in table 4.

In this chapter I introduced a grammatical framework for a fragment of syntax, motivated by the poverty of stimulus argument. All necessary principles and parameters were described in order to present a linguistic 3-parameter space (moreover, I introduced a fourth parameter, which will be of interest later). I showed how different grammars can be compared by defining a distance measure on the parameter space. Finally, the rules and parameters were used to derive all possible sentence patterns for each of the eight grammars in the parameter space.

The problem of how a learner finds the correct target grammar, i.e., the correct parameter setting, given only positive evidence, is captured by the *triggering learning algorithm*, which will be accounted for in the part to follow.

3 Triggering learning algorithm

3.1 The algorithm

The preceding sections treated the argumentation for a parametric framework in order to explain language acquisition. To sum up, in chapter 2.1.1 I introduced Gold's learnability model. I showed how Gold's theorem about the learnability of formal languages can serve as a motivation for a restriction of the grammatical search space (solving LPLA) in terms of parameters and principles, which in turn are part of the LAD. Then, I constituted a fragment of UG using a restricted system of phrase structure principles (mainly X-bar theory) and four parameters. Finally, a 3-parameter linguistic space was shown together with its language extensions.

As Briscoe states, cited on page 15, for an exhaustive formulation of the LAD we need (i) a grammatical framework containing parameters and principles, (ii) a learning algorithm to update the parameter settings, and (iii) an initial parameter setting. That is, I have already accounted for (i), but (ii) and (iii) are still to describe.

In this chapter, I will make up formulations and formalizations for a learning algorithm—namely the TLA—as well as for initial parameter settings. However, in order to entirely model the TLA, which is motivated and described in the first two sections, I will need to explain some more mathematical concepts from statistics and algebra in 3.1.3. Finally, I will give a modeling of the TLA as well as a definition of learnability including initial parameter settings. Both, the algorithm and initial parameter settings will be worked on later on in this thesis, especially in 3.2 and 4.1, respectively.

In contrast to mainly evolutionary theoretically motivated models (that is, mathematical biology), as in [Nowak 2001], [Yang 2000], or [Yang 2004], the TLA model of Niyogi and Berwick (first presented in [NB 1996]) explicitly incorporates linguistic concepts such as triggers and psycholinguistically motivated constraints, which are only implicit in the first models. Hence, language acquisition theory benefits from such an intuitive approach, since psycholinguistic knowledge about parameter settings or primary linguistic data (PLD) thus can be implemented directly in the model. In the latter part of this thesis I will show how such an implementation for both cases—parameter settings and PLD—can be worked out, to solve algorithm specific problems (so-called *local maxima*, mentioned in 3.2).

3.1.1 Triggers and constraints

Essential for the formulation of the TLA are the concept of *triggers* as well as specific *constraints*, which ensure that the algorithm behaves psychologically naturally. Both concepts are to be explained in this section.

In general, a trigger is meant to be a sentence from the target grammar, which is not analyzable (or computable, or acceptable) under the present hypothetical grammar. It is important to note that this can have two different meanings, which are captured in (23) and (24).

(23) A trigger is a sentence s from the target language L_T, which is not contained in any other possible language, i.e., $s \notin \bigcup_{i \neq T} \mathcal{L}(G_i)$.

(24) A trigger is a sentence s from the target language L_T, which is not contained in the present hypothetical language, i.e., $s \notin \mathcal{L}(G_h)$.

It is clear that if a sentence is only contained in the target language—as in (23)—, then it is also not contained in the hypothetical language—as in (24). Thus, the first notion of a trigger is logically stronger than the second one.

GW refine these definitions by including parameters. They differentiate between two different types of triggers [GW 1994, 409]:

(25) *"A* global trigger *for value v of parameter P_i, $P_i(v)$, is a sentence S from the target grammar L such that S is grammatical if and only if the value for P_i is v, no matter what the values for parameters other than P_i are."*

(26) *"Given values for all parameters but one, parameter P_i, a* local trigger *for value v of parameter P_i, $P_i(v)$, is a sentence S from the target grammar L such that S is grammatical if and only if the value for P_i is v."*

Hence, sentences in the sense of (23) are *global triggers*: A global trigger causes the learner to set a certain parameter such that the parameter then is set correctly not regarding any other parameter values.

In contrast, a *local trigger* causes the learner not necessarily to hypothesize the target grammar setting. In fact, the triggered parameter value not even has to be as in the target grammar. This is, because the sentence could be analyzable due to another parameter setting, different from that in the target grammar. Sentences described in (24) then are local triggers.

For matters of notational consistency, I will restate the definitions for global and local triggers below:

(27) A sentence $s \in L_T$ is a *global trigger* for value $v \in \{0, 1\}$ for parameter i if for all possible parameter settings $(p_1 ... p_i ... p_n)$ holds that $s \in \mathcal{L}(p_1 ... p_i ... p_n)$ iff $p_i = v$.

(28) Let $(p_1 ... p_i ... p_n)$ be a parameter setting. A sentence $s \in L_T$ is a *local trigger* for value $v \in \{0, 1\}$ for parameter x if it holds that $s \in \mathcal{L}(p_1 ... p_i ... p_n)$ iff $p_i = v$.

Note, that the difference between global and local triggers is just the position of a quantifier: Global triggers treat *all* parameter settings, whereas local triggers only have influence on one certain parameter setting.[9] Hence, it is clear that s being a global trigger immediately entails that s is a local trigger.

Remark that, if one language is a subset of another one, it is impossible to change from the superset grammar to the subset grammar, no matter whether there exist triggers—global oder local. This is, because every sentence from the subset language is also grammatical in the superset languages. Note, that this

[9] The definitions can be reformalized further in order to make the difference clear:

1. A value v for parameter i can be *triggered globally* if $\exists s \in L_T \forall (p_1 ... p_i ... p_n) : s \in \mathcal{L}(p_1 ... p_i ... p_n) \leftrightarrow p_i = v$. That is, there is one sentence, which is powerful enough to cause all parameter settings to change a certain value.
2. A value v for parameter i can be *triggered locally* if $\forall (p_1 ... p_i ... p_n) \exists s \in L_T : s \in \mathcal{L}(p_1 ... p_i ... p_n) \leftrightarrow p_i = v$. Note, however, that for binary-valued parameters the universal quantifier ranges only over one alternate parameter setting, thus being redundant.

fact (which is often referred to as *subset* or *superset problem*) is independent of the notion of a parameter—it also captures set-theoretic learnability situations as in (23) and (24). In particular, it is independent of the constraints, which shall be covered now.

Two important constraints are implemented in the TLA: the *single value constraint* and the *greediness constraint*. Both constraints are psycholinguistically motivated [GW 1994, 411]:

(29) "The Single Value Constraint

Assume that sequence $\{h_0, h_1, ..., h_n\}$ is the successive series of hypotheses proposed by the learner, where h_0 is the initial hypothesis and h_n is the target grammar. Then h_i differs from h_{i-1} by the value of at most one parameter for $i > 0$."

This constraint ensures that the learner can only move to adjacent grammars, meaning that two neighboring grammars differ in the setting of only one parameter. Following GW, there are three motivations for this constraint [GW 1994, 441f]:

1. During language acquisition, children do not vary widely in their linguistic behavior from day to day.
2. Restricting guesses to adjacent grammars the limited resources for language acquisition are represented.
3. A parameter setting should be *"tied to an input in a natural way"* [GW 1994, 442].

The greediness constraint in turn assures that only those parameters are changed, which allow to analyze the input sentence [GW 1994, 411]:

(30) "The Greediness Constraint

Upon encountering an input sentence that cannot be analyzed with the current parameter settings (i.e., is ungrammatical), the language learner will adopt a new set of parameter settings only if they allow the unanalyzable input to be syntactically analyzed."

In other words, constraint (30) embodies the hypothesis that the learner does not change his linguistic behavior if not necessary. Similar to the single value constraint also (30) accounts for the points 1 to 3 above. The learner behaves conservative in that he does not change his grammar unnecessarily, which also represents his limited resources. Finally, linguistic naturalness is ensured by prohibiting inappropriate parameter changes.

The triggers and constraints now will be included—though not explicitly mentioned—in the TLA.

3.1.2 Formulating the TLA

The original formulation of the TLA is given as follows [GW 1994, 409]—I will present an equivalent but more clearly arranged account below in (32):

(31) "The Triggering Learning Algorithm (TLA)

Given an initial set of values for n binary-valued parameters, the learner attempts to syntactically analyze an incoming sentence S. If S can be successfully analyzed, then the learner's hypothesis regarding the target grammar is left unchanged. If, however, the learner cannot analyze S, then the learner uniformly selects a parameter P (with probability $1/n$ *for each parameter), changes the value associated with P, and tries to reprocess S using the new parameter value. If the analysis is now possible, then the parameter value change is adopted. Otherwise, the original parameter value is retained."*

Restated in four steps the algorithm can be formulated in the following way (cf. [Niyogi 2006, 107f]):

(32) TLA

1. Start at some random point G_0 in the linguistic n-parameter space $\mathcal{G}^n$.
2. Receive a random example sentence $s \in L_T$.
3. If $s \in \mathcal{L}(G_h)$, that is, if s is grammatical in the current hypothetical grammar, go back to step 2. Otherwise, continue.
4. Select a single parameter at random with probability $\frac{1}{n}$. Change its value ($0 \mapsto 1$, $1 \mapsto 0$) iff $s \in \mathcal{L}(G_{h'})$, where $G_{h'}$ denotes the altered hypothetical grammar. Go to step 2.

Although the algorithm is clear and straight, several remarks have to be made:

1. The triggering properties of s are seperated into three parts: The property of s being an element of the target language is apparent in step 2, the property that triggers are ungrammatical in the hypothetical language is apparent in step 3, and step 4 contains the latter part of the definition of a trigger, namely that s is grammatical in the new hypothetical grammar. Moreover, it can be seen that the kind of triggers in the TLA is the local one. This is, because global triggers very often do not exist [GW 1994].

2. The single value constraint appears in step 4, where only one parameter is selected and potentially changed befor the next example is accounted for in step 2.

3. According to GW, the TLA is error-driven. That is, if the learner can analyze the input sentence, there is no need for him to change any of the parameter values. This is due to the greediness constraint and becomes manifest in the *iff*-condition in step 4.

4. The TLA as described here only relies on positive evidence. Any negative evidence, if it is assumed to be relevant for language acquisition, is not captured under this approach. This is due to the fact that example sentences are drawn from the target language in step 2.

5. The algorithm is of a type that is referred to as being memoryless, meaning that the decision to change a parameter value does not depend on previously selected grammars. This property is formalized in the selection of a random parameter in step 4.

6. GW mention that using the TLA it is possible to veer away from the target grammar, if only *"the input sentence is not analyzable in the current grammar, but is analyzable in the grammar that results by changing the value of a parameter that is currently set correctly"* [GW 1994, 410].

Given formulation (31), GW afford a proof for the *learnability* or *convergence* of the TLA, which is based on the assumption, that there exists triggering data for every two neighboring grammars. For the time being, let us ignore the fact that triggering data does *not* always exist, this will be handled in the next chapter.

Convergence here means that the learner guesses the correct grammar, namely the target grammar, and stays with this guess. As can be seen it is an implicit consequence of the the formulation of the TLA, that the algorithm does not halt: In the case of guessing the target grammar by fixing the correct parameter setting, the learner will go from step 3 to step 2 infinitely many times. It should be remarked, that this is due to the greediness constraint.

Before we continue, let us pause at this point. Looking back to Gold's learnability model, the question arises whether there are improvements in the TLA-approach. This objection is justified for multiple reasons: Analogously to Gold's model, the TLA-learner *learns* a certain language, if he guesses at some time t, that is, after some example sentence, the target grammar and remains there. This means that both approaches rely on convergence for a definition of learnability. Then, Gordons [Gordon 1990] explanatory notes cited on page 7 are also legitimate for the TLA:

1. It may be that the TLA-learner identifies the target grammar with his first guess. This is obviously not consistent with natural language learning situations. To make things worse, the target grammar can be "identified" by chance at step 1 in the algorithm without even getting any input sentence.

2. To guarantee learnability, i.e., convergence, the learner has to hear infinitely many input sentences. It is clear that this cannot happen, especially not within a period of several years. Or, the learner identifies the target grammar by mischance after a too large period of time.

Due to the facts that children acquire languages neither immediately, nor after, say, one century, we might be tempted to abandon this—although psycholinguistically motivated, nevertheless deficient—approach.

But, the TLA provides a significantly more natural approach to the problem of language learning. More precisely, the path through the space of grammars until finally reaching the target grammar is much more structured and linguistically natural than Gold's learnability model. This is because of structuring the search space in a linguistically appropriate way, as because of the constraints in the algorithm.

What is at least of the same importance, the TLA allows us to make statistical considerations about the learnability of languages. Let consider the following to invalidate the first objection: The probability p_n to guess the target grammar in an n-parameter space at step 1 in the algorithm is 1 divided by the number of grammars in the search space. For three parameters, we thus get a probability of $p_3 = \frac{1}{2^3} = \frac{1}{8} = 0.125 = 12.5\,\%$. For four parameters, we obtain $p_4 = \frac{1}{16} = 6.25\,\%$. If 20 parameters are assumed to constitute all natural languages, we get about $p_{20} = 0.0001\,\%$. This somehow weakens *objection 1.*

To argue against *objection 2* by calculating the probability of guessing the correct grammar after a certain time, is much trickier, but, however, solvable. The answer to this problem—in particular within a 3-parameter space—will be the topic of the rest of part 3 of this thesis. Though, before I can state a mathematical modeling of the TLA in 3.1.4, I need to introduce some necessary mathematical tools. This, then, leads us over to the subsequent section.

3.1.3 Mathematical framework – Markov chains

TLA-learning situations can be fully modeled by stochastic processes named *Markov processes* or *Markov chains*. The mathematical tools required to define the process, which are presented in this section, are mainly from probability theory, linear algebra, and fundamental graph theory. First, I will briefly review basic concepts of probability theory, then I will introduce matrices and matrix multiplication. Finally, I will state the complete graph theoretic definition of a Markov chain and its representation as transition matrix in order to—theoretically—solve the problem given at the end of the previous section.

As a motivation, consider the following simplified language learning situation: The linguistic search space $\mathcal{G}$ consists of three grammars G_1, G_2, G_3. Furthermore, by empirical investigation, we have figured out that we have the following *transition probabilities*, where $\mathbb{P}(G_i \to G_j)$ denotes the probability that the learner changes to grammar G_j if hypothesizing grammar G_i at present.

(33)

Transition probabilities		
$\mathbb{P}(G_1 \to G_1) = 1$	$\mathbb{P}(G_1 \to G_2) = 0$	$\mathbb{P}(G_1 \to G_3) = 0$
$\mathbb{P}(G_2 \to G_1) = 0.25$	$\mathbb{P}(G_2 \to G_2) = 0.5$	$\mathbb{P}(G_2 \to G_3) = 0.25$
$\mathbb{P}(G_3 \to G_1) = 0$	$\mathbb{P}(G_3 \to G_2) = 0.25$	$\mathbb{P}(G_3 \to G_3) = 0.75$

That is, if the learner currently hypothesizes grammar G_3, in 25 % of all cases he will decide to move to G_2 (maybe because of triggering sentences). In the remaining 75 % he will stay in G_3, and he never goes to G_1 from this point.

There are several important facts about the table above. First, we call the grammars G_1, G_2, G_3 *states* or *points*. Second, for each state we have exactly three transition probabilities, one for each state in the search space, and all transition probabilities are in the closed real interval $[0, 1]$. Third, if for each state all three transition probabilies are totaled, we always get a sum of exactly 1. We will refer to this as *probability distribution* (see glossary for a more detailed explanation).

Consider the following two rules of probability calculation, where A and B are events [Bronstein 2005, 772f]:

(34) $\mathbb{P}(A \textit{ or } B) = \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A and B are *mutually exclusive*, that is, if both events cannot occur at the same time.

(35) $\mathbb{P}(A \textit{ and } B) = \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$ if A and B are *independend*, that is, the occurence of one event does not affect the other one.

Now, let us calculate some probabilities: First, we want to know how likely it is that the learner moves to G_1 or G_3 if starting in G_2. Since he cannot move to both grammars from this point at the same time, both events are *mutually exclusive*. Thus,

(36) $\mathbb{P}(G_2 \rightarrow G_1 \, or \, G_2 \rightarrow G_3) = \mathbb{P}(G_2 \rightarrow G_1) + \mathbb{P}(G_2 \rightarrow G_3) = 0.25 + 0.25 = 0.5 = 50\,\%$.

Note that as mentioned above all outgoing movements sum up to 1.

Second, think of the situation where the learner starts in G_2, moves to G_3 and moves back to G_2 again. For the second move it makes no difference, how frequently a move from G_2 to G_3 takes place, vice versa. Hence, both events are *independent*. Then we end up with

(37) $\mathbb{P}(G_2 \rightarrow G_3 \, and \, G_3 \rightarrow G_2) = \mathbb{P}(G_2 \rightarrow G_3) \cdot \mathbb{P}(G_3 \rightarrow G_2) = 0.25 \cdot 0.25 = 0.0625 = 6.25\,\%$.

Third, we are interested in the question, how likely it is that the learner hypothesizes G_2 after two moves, if he originally has started in, say also, G_2. Now, we have three possibilities:

1. The learner moves from state G_2 to state G_1 and back again. Analogously to (37) we obtain $\mathbb{P}(G_2 \rightarrow G_1 and G_1 \rightarrow G_2) = 0.25 \cdot 0 = 0$. This is, because if once in G_1, the learner never moves back from there.

2. The learner stays in G_2 twice. Then, we calculate $\mathbb{P}(G_2 \rightarrow G_2 \, and \, G_2 \rightarrow G_2) = 0.5 \cdot 0.5 = 0.25$.

3. The learner moves from state G_2 to state G_3 and back again. We have already calculated this probability in (37), i.e., 0.0625.

That is, the learner choses movement 1, or 2, or 3. Since only one of them can occur at once, all events are mutually exclusive. Hence, we get

(38) $\mathbb{P}(G_2 \rightarrow G_2 \, after \, two \, steps) = 0 + 0.25 + 0.0625 = 0.3125 = 31.25\,\%$.

In the remaining 68.75 % he will end up in either G_1 or G_3. It is useful to keep this calculation in mind, as it will be important soon.

Now, after the basic probability theoretic concepts have been introduced, let us move over to *matrices* (see for instance [Howard 1998]). An $m \times n$-*matrix* T is defined as a rectangular schema of numbers with m rows and n columns. The numbers are referred to as the *elements* or *entries* of the matrix. The elements are conventionally indexed $T_{ij}$where i is the number of the row and j the number of the column. For our purpose, only square matrices such as the 3×3-matrix below are interesting.

(39)
$$T = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0.25 & 0.5 & 0.25 \\ 0 & 0.25 & 0.75 \end{pmatrix}$$

The entries of the matrix above are just the transition probabilities of the table in (33).

Matrix multiplication is defined in the following manner: To get the entry C_{ij} of the product matrix $C = A \cdot B$, we compute the pairwise products of the elements in the ith row of A and the jth colums of B and sum them up. Thus,

(40) $C_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j} + ... + A_{in}B_{nj}$.

If now looking back to the probability theoretic example above, we find that (38) was calculated *exactly* as the entries of the matrix product $T \cdot T = T^2$, the (second) *power* of matrix T. We then obtain

(41)
$$T^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0.375 & 0.3125 & 0.3125 \\ 0.0625 & 0.3125 & 0.625 \end{pmatrix}.$$

Similarly, we can calculate T^k by self-multiplying the matrix k times. Thus, the entries T^k_{ij} of the matrix are nothing but the probabilies to get from state G_i to state G_j in exactly k steps. This will lead us further to the answer of the question of the end of the preceding chapter.

Moreover, if we multiply the matrix an infinite number of times, that is if we build the limit[10] $\lim\limits_{k\to\infty} T^k$, we can calculate

(42)
$$T^\infty := \lim_{k\to\infty} T^k = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

I am now ready to formulate the definition of a Markov chain. As it is more intuitive than a probability theoretic definition as in [Feller 1950, 374] (see glossary; note that they differ, in that the following approach does not cover the initial probability distribution), I will—just as [Niyogi 2006]—define Markov chains in graph theoretic terms.

(43) *Markov chain*

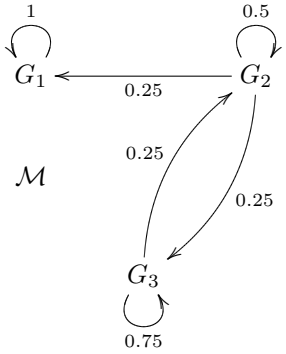
A *Markov chain* $\mathcal{M}$ is a directed, weighted, probablistic graph, i.e., an ordered pair (V, E), where V is the set of n *vertices* or *states* and E is the set of n^2 directed, weighted *edges*, such that for each state there are exactly n outgoing edges, whose weights total to 1 in each case.

In the definition *directed* means that each edge is an ordered pair of two vertices, *weighted* means that each edge is associated with a certain number, and *probablistic* denotes that the weights on the outgoing edges form a probability distribution $\mathbb{P}$.

The graph below is an example for a Markov chain. Note that zero-weighted edges are omitted.

[10] One can show that this limit always exists if the columns or rows form a probability distribution, as in our case.

(44)



The weights in $\mathcal{M}$ are just the same as the entries in T. The example illustrates that for each Markov chain we can find a corresponding matrix T with entries T_{ij} for each edge from state i to j, and vice versa. For a Markov chain $\mathcal{M}$, we define the *transition matrix* $T_{\mathcal{M}}$ to be this well-defined corresponding matrix. It is now evident that Markov chains share the same properties with transition matrices. In particular, we can calculate the transition probabilities of n movements or steps for arbitrary starting and end points.

Hereby I presented all mathematical basics necessary to state a modeling of the TLA. This, however, is done in the next section.

3.1.4 Modeling the TLA

Given the mathematical foundations described in the previous section, what exactly is necessary to fully model the language learning situation for a TLA-learner? For convenience, let the formulation be repeated here:

(45) TLA

1. Start at some random point G_0 in the linguistic n-parameter space $\mathcal{G}^n$.
2. Receive a random example sentence $s \in L_T$.
3. If $s \in \mathcal{L}(G_h)$, that is, if s is grammatical in the current hypothetical grammar, go back to step 2. Otherwise, continue.
4. Select a single parameter at random with probability $\frac{1}{n}$. Change its value $(0 \mapsto 1,\ 1 \mapsto 0)$ iff $s \in \mathcal{L}(G_{h'})$, where $G_{h'}$ denotes the altered hypothetical grammar. Go to step 2.

Looking thoroughly at each of the four steps, we can point out the subsequent demands (in notational matters I predominatly hold on to [Niyogi 2006], as well as to the already introduced notation):

1. In step 1 a random point is selected, that is, we need a probability distribution Π_0 on the space of grammars. This distribution represents a preference for certain parameter settings. By setting $\Pi_0(G) = 0$ for some G, certain parameter settings can be excluded of being a starting point. This is discussed extensively in [Hyams 1986] and [GW 1994]. I will return to it later in part 4.

2. In step 2 a random sentence of the target language is presented. Consequently, we need a probability distribution P on the sentences of L_T. More precisely, P is a probability distribution on the sentence patterns of the target grammar. Such a distribution for primary linguistic data can be empirically determined, e.g. via CHILDES data analysis.

3. Step 3 states that only in the unaccepted cases the learner will hypothesize another grammar. Formally, if q is the probability to go to another grammar, then the learner stays in the present grammar G_h with a probability of $\mathbb{P}(G_h \to G_h) = 1 - q$.

4. Several points have to be considered in step 4:

   (a) At most one parameter can be changed for each example sentence. That is, according to the single value constraint, the learner cannot move to a not-adjacent grammar. So, all Grammars, which differ from G_h in more than one parameter value, are selected with probability 0. Formally, I refer to the distance measure on parameter spaces, and thus can state that, if $d_{hamm}(G_h, G_{h'}) > 1$, then $\mathbb{P}(G_h \to G_{h'}) = 0$.

   (b) The sentences are drawn from the target language. Thus, we are only interested in sentences of a certain language $\mathcal{L}(G_i)$ that are also in $s \in L_T$. Let this corresponding set be denoted by $L_i^T := L_T \cap \mathcal{L}(G_i)$.

   (c) Parameters are selected with probability $\frac{1}{n}$, but the probability to chose an adjacent grammar $G_{h'}$ (i.e., $d_{hamm}(G_h, G_{h'}) = 1$) also depends on the quantity of sentences of the target language that are analyzable in $G_{h'}$ but not in G_h. Since, if a sentence s is not in $L_{h'}^T$, the learner would not go there due to the greediness constraint, no matter whether s was in L_h^T or not. Such a sentence is drawn with probability $P(L_{h'}^T \setminus L_h^T)$. Altogether, grammar $G_{h'}$ is selected with probability $\mathbb{P}(G_h \to G_{h'}) = \frac{1}{n} P(L_{h'}^T \setminus L_h^T)$.

   (d) In total, the probability q to move to any other grammar starting in G_h then is the sum of all probabilities to move to another grammar (cf. calculation (36) for this point), hence $q = \sum_{h \neq h'} \mathbb{P}(G_h \to G_{h'})$.

5. Then $\mathbb{P}(G_h \to G_h) = 1 - \sum_{h \neq h'} \mathbb{P}(G_h \to G_{h'})$.

Thus, in the way described above, for all grammars G_h and $G_{h'}$ we can calculate the transition probabilities $\mathbb{P}(G_h \to G_{h'})$. Hence, we can associate the TLA for a linguistic n-parameter space $\mathcal{G}^n$ with a Markov chain $\mathcal{M}$ and its corresponding transition matrix

(46)
$$T_{\mathcal{M}} = \begin{pmatrix} \mathbb{P}(G_1 \to G_1) & \mathbb{P}(G_1 \to G_2) & \cdots & \mathbb{P}(G_1 \to G_{2^n}) \\ \mathbb{P}(G_2 \to G_1) & \mathbb{P}(G_2 \to G_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbb{P}(G_{2^n} \to G_1) & \cdots & \cdots & \mathbb{P}(G_{2^n} \to G_{2^n}) \end{pmatrix}.$$

The space has 2^n grammars or states, so there are 2^{2n} transition probabilities or weights in the Markov chain, some of which equal 0 for $n > 1$.

Note that for the target language as hypothetical grammar we always obtain

(47) $\mathbb{P}(G_T \to G_i) = \frac{1}{n}P(L_i^T \setminus L_T^T) = \frac{1}{n}P(L_i^T \setminus L_T) = \frac{1}{n}P(\emptyset) = 0,$

for any $i \neq T$ since by definition every sentence in the relativized language L_i^T is also contained in the target language. Consequently, then, $\mathbb{P}(G_T \to G_T) = 1 - \sum 0 = 1$. That is, once the learner reaches the target grammar, he will chose it again with probability 1.

As Niyogi states "[t]*his model captures the dynamics of the TLA completely. We note that the learner's movement from one language hypothesis to another is driven by purely extensional considerations—that is, it is determined by set differences between language pairs"* [Niyogi 2006, 117f].

Finally, we can make up a definition for learnability in terms of Markov chains. The definition given below is a significantly reviewed version of *Definition 10* in [Niyogi 2006]. Before, we need one preparing definition stated here (see footnote for an explanation):

(48) For the initial probability distribution $\Pi_0 = (\pi_1, ..., \pi_{2^n})$ we can define $\Pi_k = \Pi_0 T_{\mathcal{M}}^k$, which is the probability distibution on the space of grammars after hearing k sentences. That is, $\Pi_k = (\pi_1(k), ..., \pi_{2^n}(k))$ is a row vektor and $\pi_i(k)$ denotes the probability of being in state G_i after hearing k sentences.[11]

(49) *Learnability*

Let $\mathcal{G}^n$ be a linguistic n-parameter space of ennumerated grammars, let $G_T \in \mathcal{G}^n$, let Π_0 be a probability distribution on $\mathcal{G}^n$ determining the starting point, let P be a probability distribution on $\mathcal{L}(G_T)$, and let $\mathcal{M}$ be a 2^n-state Markov chain on $\mathcal{G}^n$ with transition probabilities determined by P.

The language $\mathcal{L}(G_T)$ is said to be *learnable* if $\mathcal{M}$ selects G_T with probability 1 in the limit according to Π_0, that is, if for the transition matrix $T_{\mathcal{M}}$ the following holds:

$\Pi_\infty := \Pi_0 T_{\mathcal{M}}^\infty$ has only zeros as entries exept for the Tth entry, which equals 1, i.e.:

$$\Pi_\infty = (0, \ldots, 0, \underset{\substack{\uparrow \\ T\text{th}}}{1}, 0, \ldots, 0).$$

[11] To make the definition clear, for $\Pi_0 = (0.3, 0.3, 0.4)$ consider how the entry $\pi_3(1)$ is calculated for the Markov chain given in (44): The third entry of $\Pi_1 = \Pi_0 T_{\mathcal{M}}$ is

$$\pi_3(1) = 0 \cdot \pi_1 + 0.25 \cdot \pi_2 + 0.75 \cdot \pi_3 = 0 \cdot 0.3 + 0.25 \cdot 0.3 + 0.75 \cdot 0.4 = 0.375.$$

From state G_1 the learner does not move anywhere, thus $0 \cdot 0.3$. From state G_2 in 25 % of all cases the learner goes to G_3, but he only starts in G_2 with probability 0.3, so we get $0.25 \cdot 0.3$. Finally, in 40 % of all cases the learner starts in the third grammar and will stay there with probability 0.4. Since all three events are mutually exclusive (he can only start at exactly one point in the chain), the probabilities sum up to 0.375.

The calculation is analogous for $\Pi_k = \Pi_0 T_{\mathcal{M}}^k$.

This definition completes the mathematical modeling of the TLA: I discussed the space of possible grammars in parametric terms, the way they are selected during language acquisition using Markov chain theory, and at last I provided a definition for learnability in this framework. We are now up to make predictions about the learnability of natural parameter settings: Meaning, that we can state *whether* they are learnable as well as *how fast* they can be acquired.

3.2 TLA in a linguistic 3-parameter space

In the first chapter of part 3 I presented the formulation for the TLA as well as its mathematical modeling via Markov chain theory. Especially in the last two sections I mainly introduced mathematical conceps, now, it is time to reap the fruits of our labor: Transition matrices shall be used to compute identification probabilities for large numbers of example sentences. As subject of study I will resort to the linguistic 3-parameter space evaluated in 2.2.4. That is, I will work on the languages given in table 4 on page 21, which were generated by the *spec* parameter, the *comp* parameter, and the *verb second* parameter.

In the first two sections I will focus on the parameter setting for (a fragment of) German. I will compute the corresponding transition matrix and assess its learnability properties by presenting learnability curves for all possible starting grammars as well as a learnability curve which refers to an initial probability distribution. Furthermore, I will work out a sufficient learnability condition on the form of transition matrices.

In the last two sections of this chapter I will treat the parameter setting for (again, a fragment of) English. Here, I will also compute the corresponding transition matrix, revealing problematic points in the space of grammars, known as *local maxima*. These are states in the linguistic search space, where the learner can get stuck, such that he will never reach the target grammar, thus affecting learnability, as learning curves will show.

Since the problem of local maxima also concerns natural language parameter settings, there exists an attempt to find a solution for this problem in literature. Though, solving this problem is not purpose of this chapter, which basically employs illustrative computer assisted calculations. Several solutions will be focussed on in a final step in 4.

3.2.1 Computing the transition matrix

For calculating a transition matrix for the linguistic 3-parameter space, the extensions of the languages are needed. A list of all languages, generated by p_{spec}, p_{comp}, and p_{V2} is given in table 4 on page 21 or in appendix A.1.

Suppose, the fragment representing the German paremeter setting is the target grammar, that is, $G_T = G_8 = (001)$. Furthermore, we assume that the sentences in L_T are *uniformly distributed*, i.e. the probability to draw a sentence s of a subset A of L_T is $P(A) = \frac{|A|}{|L_T|}$, where $|\,.\,|$ denotes the number of elements in a set. For example, $|L_T| = 19$, since there are 19 different sentence patterns in the target language for $T = 8$.

We want now to calculate the transition probability to go from $G_1 = (110)$ to $G_2 = (111)$. It is clear, that both grammars are not adjacent to the target grammars, though the languages $\mathcal{L}(G_1)$ and $\mathcal{L}(G_2)$ share some sentence patterns with $L_T = \mathcal{L}(G_8)$. Now, the relativized languages as in 4b on page 32:

(50) $L_1^T := L_T \cap \mathcal{L}(G_1) = \{\text{ADV V S}\}$ and

(51) $L_2^T := L_T \cap \mathcal{L}(G_2) = \{\text{ADV V S, O V S, S AUX V, S V, S V O}\}$.

Grammar G_1 has the greatest distance to the target grammar, hence both corresponding languages share only one sentence pattern, ADV V S, whereas the relativized language for G_2 contains 5 elements. Then we count the sentences that are acceptable for G_2 but not for G_1, yielding

(52) $|L_2^T \setminus L_1^T| = |\{\text{O V S, S AUX V, S V, S V O}\}| = 4.$

Thus, the probability to draw a sentence of the target grammar, which is also in the set in (52), is $P(L_2^T \setminus L_1^T) = \frac{4}{19}$. But, since starting in G_1, there are 3 possible grammars to go to (the learner randomly selects one of the 3 parameters), the probability to move to G_2 is

(53) $\mathbb{P}(G_1 \to G_2) = \frac{1}{3} \cdot \frac{4}{19} = \frac{4}{57} = 0.07017... \approx 7\%.$

The calculation (50)-(53) exemplarily shows how the entries of the transition matrix are computed. Keep in mind, that for grammars, which are farther away from a given hypothetical grammar than 1 parameter change, we directly receive 0 as transition probability, due to the single value constraint. Moreover, the transition probability from the target grammar to itself equals 1 and all other outgoing weights are 0. The complete transition matrix then is

(54)
$$T_{\mathcal{M},8} = \begin{pmatrix} \frac{49}{57} & \frac{4}{57} & \frac{1}{57} & 0 & \frac{1}{19} & 0 & 0 & 0 \\ 0 & \frac{49}{57} & 0 & \frac{1}{19} & 0 & \frac{5}{57} & 0 & 0 \\ 0 & 0 & \frac{50}{57} & \frac{2}{19} & 0 & 0 & \frac{1}{57} & 0 \\ 0 & 0 & 0 & \frac{46}{57} & 0 & 0 & 0 & \frac{1}{57} \\ \frac{1}{57} & 0 & 0 & 0 & \frac{49}{57} & \frac{7}{57} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{16}{19} & 0 & \frac{3}{19} \\ 0 & 0 & \frac{2}{57} & 0 & \frac{2}{57} & 0 & \frac{35}{57} & \frac{6}{19} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where the number in the index refers to the target grammar (see appendix A.3 for the codes that were used to computationally calculate the matrix). Note that the only value of 1 is in $\mathbb{P}(G_8 \to G_8)$, that is, for each starting point there is a certain chance for the learner to reach the target grammar. As we will see, this is not always the case. Furthermore, there are at most four non-zero entries in each row: one for each adjacent grammar plus one for staying in the hypothetical grammar. If there are less than four entries, the learner would not chose some adjacent grammars because of the greediness contraint.

By computing the limit of $T_{\mathcal{M},8}^k$ for $k \to \infty$, we obtain the matrix

(55)
$$T_{\mathcal{M},8}^{\infty} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

That is, no matter where the learner starts, as the number of given examples goes to infinity, he will converge to the target grammar G_8. Moreover, the initial probability distribution Π_0 has no effect on this outcome. Regardless of how the grammars are weighted at the beginning of the learning process, one always obtains

$$\Pi_\infty = \Pi_0 \cdot T^\infty_{\mathcal{M},T} = (0, \ldots, 0, \underset{\substack{\uparrow \\ T\text{th position}}}{\pi_1 \cdot 1 + \ldots + \pi_{2^n} \cdot 1}, 0, \ldots, 0) =$$

$$= [\Pi_0 \text{ is a probability distribution}] = (0, \ldots, 0, 1, 0, \ldots, 0)$$

for limit transition matrices $T^\infty_{\mathcal{M},T}$, such that the Tth column consists of ones only, and arbitrary probability distributions Π_0. Hence, according to the definition of learnability for the TLA we obtain the following learnability condition:

(56) If for a target grammar G_T the limit transition matrix $T^\infty_{\mathcal{M},T}$ has only ones in the Tth column, the language L_T is learnable.

So, we have a *sufficient* condition for learnability within our framework, a condition, which can be computed with some theory and a little help from a computational system. That is, for each parameter based fragment of UG we can evaluate, (i) whether a certain language is learnable by the fact above, and (ii) whether or not a certain language is learnable, if a probability distribution on the space of grammars is given, by calculating Π_∞. The subsequent section will account for the question, how many example sentences it takes to identify a target grammar with a particular probability.

3.2.2 Convergence curves for learnable grammars

The transition matrix $T_{\mathcal{M},8}$ for target grammar $G_8 = (001)$ computed in the preceding section now will be investigated further. In particular, I will study the transition probabilities for larger numbers of example sentences. As we have seen in 3.1.3, the entry in the ith row and jth column of a matrix T^k, where T is a transition matrix and $k > 0$, just is the probability to go from state i to state j in exactly k steps. For notational reasons, I will write $T^k_{\mathcal{M},T}[i,j]$ for the entry ij of $T^k_{\mathcal{M},T}$.

Then, by computing $T^k_{\mathcal{M},8}$ for some $k > 0$, one can calculate the probability of reaching state T when starting in state i, where $1 \leq i \leq 8$, in k steps. Or, to put it in another way, one can calculate the probability of acquiring the target grammar after hearing k example sentences from L_T, when the learner has started in one of the grammars G_1 to G_8.

That is, we compute $T^k_{\mathcal{M},8}[i,8]$ (or $T^k_{\mathcal{M},T}[i,T]$ in general). If this is done for, say, $1 \leq k \leq 50$, and plotted for all possible starting grammars, we result in the identification probability graph in figure 2.

The horizontal axis *nmbr* denotes the number of given examples, while the vertical axis *prob* measures the probability to identify the target grammar after *nmbr* examples given. The starting grammars are denoted by their indices.

It can be seen, that after 50 example sentences the learner acquires the target grammar with a probability of nearly 100 % for each target grammar. Furthermore, the convergence curve for the target grammar is the constant function $prob = 1$. The other curves can be divided into two classes: The

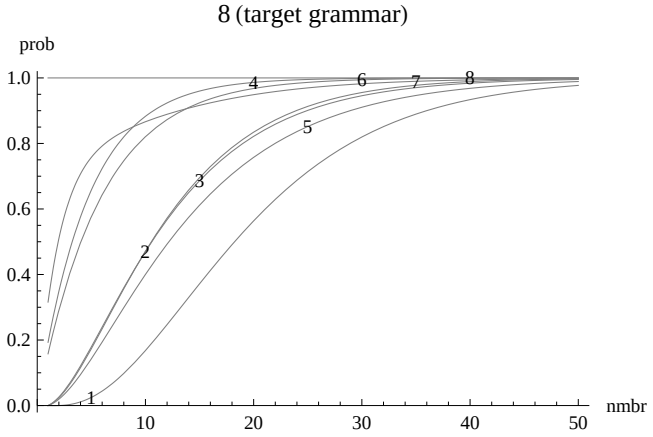


Figure 2: *Identification probabilities for target grammar G_8 and 50 example sentences.*

concave curves 4, 6, 7 represent exaclty those grammars, which are adjacent to the target grammar, while the sigmoidal curves 1, 2, 3, 5 represent the non-adjacent grammars. In particular, the most isolated convergence curve is the one for $G_1 = (110)$, which is 3 steps away from the target grammar.

It is interesting that the curve for grammar G_7 (fragment of Hindi) first is strongly increasing but flattening afterwards. This is, because unlike all grammars, but the anyway almost isolated grammar G_1, this grammar shares G_8-sentence patterns with two other non-target grammars, namely G_3 and G_5, and due to the fact that the other two adjacent grammars (G_4 and G_6) do not share any G_8-sentence patterns with another grammar.

Now, assume that the starting grammars are uniformly distributed, that is, to select a grammar G_i the learner has a probability $\Pi_0(G_i) = \frac{1}{8}$, thus $\Pi_0 = (\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. Then, we can compute Π_k for $1 \leq k \leq 50$, and observe the 8^{th} coordinate, whereby we get the convergence curve for the acquisition of G_8 according to the probability distribution Π_0, shown in figure 3.

What the graph shows is, that the learner converges to G_8 with probability 1, and that it takes about 50 example sentences to acquire the language. Certainly, this is not in accordance with any natural language learning situation, but we have to bear in mind, that we are examining a rather restricted grammatical framework here (only three parameters out of, maybe more than, 20), and that the size of the parameter space increases exponentially.

However, it is not always the case that grammars can be identified with probability 1 like the one observed in this section. I will account for TLA-unlearnable grammars in the following two sections.

3.2.3 Local maxima

In the previous section I analyzed the learnability properties of grammar G_8. Now suppose, the target language is $G_5 = (010)$, a fragment of English grammar.

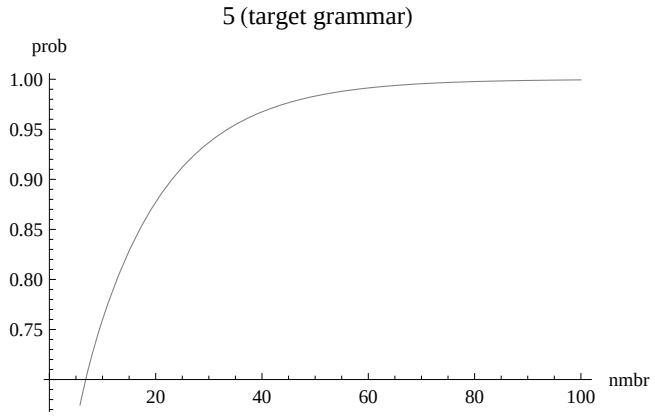


Figure 3: *Convergence curve for target grammar G_8 according to a uniform distribution Π_0.*

If we compute the transition matrix for the new target grammar in the already described way, we obtain

(57)
$$T_{\mathcal{M},5} = \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{31}{36} & \frac{1}{12} & 0 & 0 & \frac{1}{18} & 0 \\ 0 & \frac{1}{12} & 0 & \frac{11}{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{5}{6} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{5}{18} & 0 & \frac{2}{3} & \frac{1}{18} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & \frac{1}{36} & \frac{8}{9} \end{pmatrix}.$$

As expected, in the fifth row there is a one in the fifth column—the target grammar has index 5—and zeros elsewhere. But, in contrast to $T_{\mathcal{M},8}$ the matrix above contains an additional row with the same property: In the second row $T_{\mathcal{M},5}[2,2] = 1$, and all other entries are set 0. Graph theoretically speaking, there are no outgoing edges, which are weighted greater than 0, exept one going back to G_2, just as for the target grammar. We will refer to such points as *local maxima.*[12]

For illustration, see figure 4 on the next page, where the transition matrix is given in their original Markov chain form. The graph shows that there are three more problematic grammars:

[12]Note that there is a terminological difference to how local maxima are defined in [GW 1994]. GW define local maxima via connectedness to the target language. In this case, both G_2 and G_4 are local maxima, since from these states there is no path to the target. I think that, since the learner will not stay G_2 in the limit, such a definition is not intuitive.

[NB 1994] differenciate between *sink* and *not sink*, but this says nothing about whether there is a chance to go to the target or not (i.e., the difference between G_4 and G_1). So, I introduce the notions of *dead ends* and *potential dead ends*. However, this terminological question is only a matter of taste.

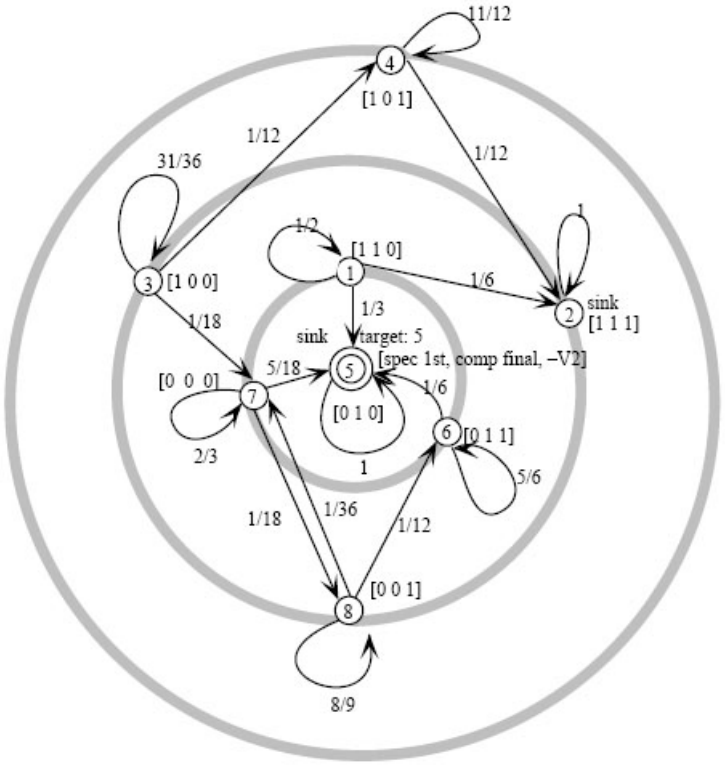


Figure 4: *Linguistic 3-parameter space with target grammar $T = 5$ modeled as Markov chain, taken from* [Niyogi 2006, 115]. *The weights on the edges are given as fractures, and the eight grammars G_1 to G_8 are given by their parameter settings and denoted by their indices. The concentric circles represent the increasing Hamming distance from each grammar to the target grammar. Grammars on a circle are two steps away from each other.*

1. G_4 only leads to the local maximum and to no other grammar, except back to itself. Let us call grammars with this property *dead ends*.

2. G_1 leads to the target grammar but also to the local maximum with probability $\frac{1}{6}$. I will refer to such grammars as *potential dead ends*.

3. G_5, although not directly leading to the local maximum (it is too Hamming-far away), leads to G_4 with probability $\frac{1}{12}$, which is a problematic state itself. Thus, G_5 also is a potential dead end.

To investigate the learnability of G_5 let us have a look at the limit transition matrix:

(58)
$$T^{\infty}_{\mathcal{M},5} = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{3}{5} & 0 & 0 & \frac{2}{5} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Thus, starting in G_1 the learner has a chance of about 67 % to reach the target, in the case of G_3 it is 40 %, and starting in G_2 the probability to identify the target grammar is 0, as expected. For starting points G_6 to G_8 there are no learnability problems. Note that the local maximum G_2 has parameter setting (111), and that $d_{hamm}(G_2, G_T) = 2$. Thus, a local maximum has not necessarily to be far away from the target grammar.

The fact that there are three more problematic states has been overlooked by GW, as they only rely on connectedness to determine local maxima (see footnote in [Niyogi 2006, 130]).

However, learnability is not only determined by the form of the transition matrix, but also by the initial probability distribution on the space of grammars. Thus, if we put $\Pi_0 = (0, 0, 0, 0, \pi_5, \pi_6, \pi_7, \pi_8)$, i.e., all problematic grammars are excluded as starting points and the rest is distributed equally, we could still obtain learnability. This topic will be discussed in 4.1.2.

So we see that we cannot formulate a *necessary* condition for learnability depending on the form of the transition matrix, as a counterpart to (56). Let us continue our investigation by considering some convergence curves for identification probabilities.

3.2.4 Convergence curves for unlearnable grammars

In section 3.2.2 I examined convergence curves for grammars that are learnable for any initial parameter setting. As we have seen, for target grammar G_5 there exist some problematic states. The graphs in this section will shed more light on the convergence behavior for potentially unlearnable grammars.

In figure 5 the curves can be divided into three classes: 6, 7 and 8 all converge to the target grammar (G_5, again a constant function), where the first two are C-shaped, since both are neighboring grammars of G_5. The latter one, 8, is S

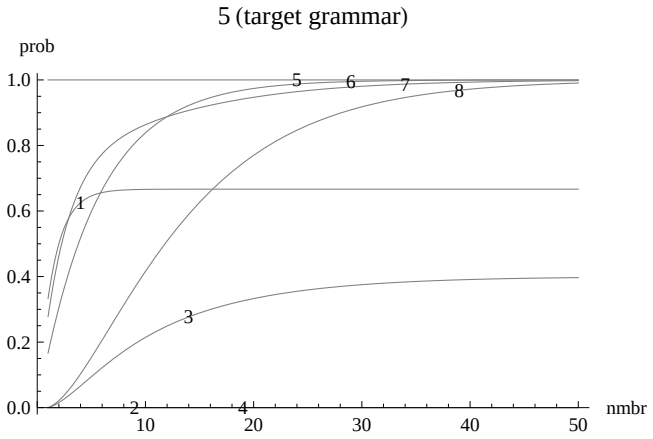


Figure 5: *Identification probabilities for target grammar G_5 and 50 example sentences.*

shaped, as G_8 is two steps away from the target. Where G_6 and G_7 have direct access to the target, coming from G_8 the learner must take a detour via G_6 or G_7.

The curves 2 and 4 are contstantly zero: This is not surprising, as both of them have no access to any other grammar. More interesting are 1 and 3, which converge to (approximately) 0.67 and 0.4, respectively. While the first one abruptly flattens after a rise, the second one converges rather slowly.

As in the graph in 3.2.2 all curves converge to their limit after at most 50 example sentences. This means that the learner recognizes rather quickly when he gets stuck in a local maximum. Let us now observe how learnability according to an initial probability distribution is affected by the presence of local maxima.

Let again $\Pi_0 = (\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$, i.e., a uniform distribution. Then we can compute the limit yielding $\Pi_\infty = (0.3667, 0, 0, 0.6333, 0, 0, 0)$ and the convergence curve is as given in figure 6. The curve converges to 0.6333 after about 50 example sentences. In the remaining 36.67 % of all cases the learner will get stuck in the local maximum.

However, this outcome can easily be influenced by stating the distribution in another way. If the values for π_1 to π_4 are decreased (or probably set to 0), and the other values are increased, then the learnability rate will increase, too (probably to 1). This, then would represent preset parameter values, a widely discussed topic in psycholinguistic learnability theory. Since we cannot ad hoc set Π_0 as it would be pleasant, we are in need for solid justifications for preset parameters. Otherwise, a uniform distribution as above would have to be assumed. See for instance [Hyams 1986] for several proposals.

In this chapter I presented an analysis of the behavior of the TLA in a linguistic 3-parameter space, consisting of eight grammars. I computed transition matrices for two of the eight grammars, of which the first one—namely G_8—was unproblematic. I presented identification probability curves for G_8

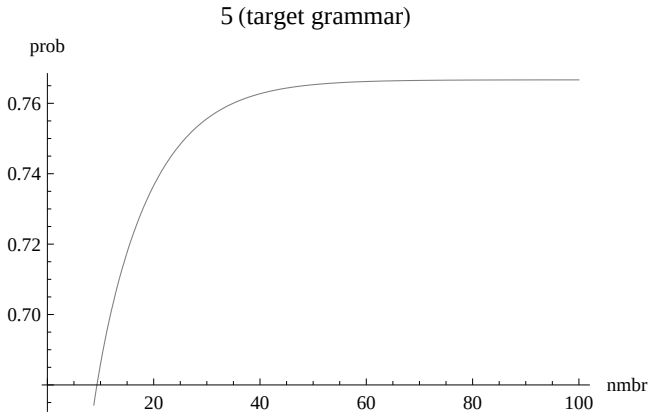


Figure 6: *Convergence curve for target grammar G_5 according to a uniform distribution Π_0.*

and showed, how long it takes the learner to identify the target with a high probability. Furthermore, I presented a sufficient condition for learnability.

The transition matrix for the second grammar which has been investigated (target G_5), however, leads to problematic states called *local maxima*, where the learner cannot escape. I showed, how learnability curves for such a grammar look like and how learnability crucially depends on the initial probability distribution.

There are several proposals how the problem of local maxima could be solved, which I will examine in the final part of this thesis.

4 Solutions for the local maxima problem

4.1 GW's analysis and NB's remark

The investigation of the TLA shows that there exist problematic states for certain target grammars. In particular, there exist local maxima, where the learner gets stuck once reached in the learning procedure. Obviously, if the TLA is seen as a psycholinguistically plausible model of language acquisition, the fact that the fragment of an existing language is only acquired with a probability of 0.63 is contradictory.

Thus, GW propose several solutions for the problem of local maxima, some of which are unfortunately unsuitable: Since [GW 1994] rely only on connectedness in one direction when classifying problematic states, they miss exactly those problematic states, which are connected with the target as well as with local maxima, as [NB 1994] demonstrated. In particular, the Markov learning model developed by NB allows to make assertions of how likely it is to converge to a local maximum, instead of making pure statements about learnability. Remember Gordon's comments on learnability models on page 7.

In the first section, I will investigate variants of the TLA where the single value constraint and the greediness constraint are omitted. Niyogi often has pointed out the flexibility of the Markov formulation, consequently I will show how to model the variants. Convergence curves are shown for quantitative comparisons.

In section 4.1.2 I will provide a modeling for preset parameter values and parameter orderings, which has not been covered by NB's investigations, as far as I know. The model shows, that the local maxima problem cannot be avoided, but decreased.

Section 4.1.3 treats the assumption that there might exist (formally) unlearnable grammars, immediately abandoned by GW. As in current theories of language change not acquiring a language plays a central role, I will focus on the question, whether or not local maxima are of relevance for diachronic syntax.

Finally, the remaining solutions proposed by GW are examined. The fact that none of the solutions in this chapter can be considered satisfactory for different reasons, will lead us to another assumption focussed on in the subsequent chapter.

4.1.1 Relaxing constraints

The first attempt to solve the problem of local maxima is to reformulate the algorithm. The TLA is crucially based on the concepts of two constraints, namely the single value constraint as well as the greediness constraint. Both constraints are thought to ensure a linguistically natural behavior of the learning algorithm, but they are also responsible for the existence of local maxima. In this section I will focus on the behavior of variants of the TLA without the single value constraint or the greediness constraint, respectively.

Let us at first handle the *single value constraint*, which shall be restated here for convenience:

(59) "The Single Value Constraint

Assume that sequence $\{h_0, h_1, ..., h_n\}$ *is the successive series of hypotheses proposed by the learner, where* h_0 *is the initial hypothesis and*

> *h_n is the target grammar. Then h_i differs from h_{i-1} by the value of at most one parameter for $i > 0$."* [GW 1994, 411]

In the mathematical modeling described in the previous chapter the single value constraint causes that from one state to the other there are positive transition probabilites only if the states are neighbored, that is, if the Hamming distance between both states is 1. This also means that for some grammars the target grammar is not directly accessible, such that the learner has to take a detour. Thus, even if the learner would be triggered out of some distant hypothetical grammar to the target (due to the greediness constraint), he is not allowed to directly go there. As a consequence, then, if there exists no triggering data for all neighboring grammars, the distant grammar becomes a local maximum.

If now this constraint is removed, there can be no local maxima, since then, nothing prevents the learner from jumping over the grammar, which is blocking the move to the target grammar. The downside—however—is that then the TLA lacks psycholinguistical adequacy: The learner could move from one hypothetical grammar to an arbitrarily distant one, which is not necessarily closer to the target than the source. This means that also linguistic behavior can vastly change within a short period of time, and "[t]*hus, abandoning the Single Value Constraint without an alternative constraint that retains conservatism is undesirable*" [GW 1994, 442].

How does removing this constraint affect the mathematical modeling? Since in this case for the learner hyothesizing a grammar G_h not only the neighbored grammars are potential new hypothetical grammars, the condition "*if* $d_{hamm}(G_h, G_{h'}) > 1$, *then* $\mathbb{P}(G_h \to G_{h'}) = 0$" simply is omitted in the computation. Thus, in the Markov chain there are up to $2^n - 1$ outgoing positively weighted arcs. For illustration, consider the transition matrix $T_{\mathcal{M}_{no\,svc},5}$ below, which is calculated for target grammar G_5.

(60)

$$T_{\mathcal{M}_{no\,svc},5} = \begin{pmatrix} \frac{1}{9} & \frac{1}{6} & 0 & \frac{1}{12} & \frac{1}{3} & \frac{1}{6} & \frac{1}{18} & \frac{1}{12} \\ 0 & \frac{29}{36} & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{36} & 0 \\ 0 & \frac{1}{6} & \frac{1}{9} & \frac{1}{12} & \frac{1}{3} & \frac{1}{6} & \frac{1}{18} & \frac{1}{12} \\ 0 & \frac{1}{12} & 0 & \frac{5}{9} & \frac{1}{4} & \frac{1}{12} & \frac{1}{36} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{29}{36} & \frac{1}{36} & 0 \\ 0 & \frac{5}{36} & 0 & \frac{1}{18} & \frac{5}{18} & \frac{5}{36} & \frac{1}{3} & \frac{1}{18} \\ 0 & \frac{1}{12} & 0 & 0 & \frac{1}{4} & \frac{1}{12} & \frac{1}{36} & \frac{5}{9} \end{pmatrix}$$

Exept for the fifth row, all zeros are only because of the fact, that the respective languages share no sentence patterns relativized to the target. Note that for each grammar the transition probabilities to neighboring grammars are the same as in (57).

Originally G_5 is a local maxima grammar with a local maximum in G_2. It is worth to remark that, although the probability to stay in G_2 is comparatively high, it is not a local maximum anymore. Figure 7 shows the convergence curves for $T_{\mathcal{M}_{no\,svc},5}$. As can be seen, although the algorithm lacks psycholinguistical adequacy, the learner converges to the target much faster without the single value constraint. This agrees with Niyogi's analysis [Niyogi 2006, ch. 4.2.1]. It should be mentioned that Niyogi studied learnable grammars in this case.

The *greediness constraint* is motivated by the idea that a learner leaves a hypothetical grammar only if the change is useful, meaning that he is able to analyze a sample sentence using the new hypothetical grammar. If for all examples the neighboring grammars are of no avail, then the learner will not leave the present state and thus get stuck in a local maximum.

(61) "The Greediness Constraint

> *Upon encountering an input sentence that cannot be analyzed with the current parameter settings (i.e., is ungrammatical), the language learner will adopt a new set of parameter settings only if they allow the unanalyzable input to be syntactically analyzed."* [GW 1994, 411]

If we remove the greediness constraint, the criterion for decision is inverted: At first, the learner tests whether a sentence is grammatical under the hypothetical grammar. If this is the case, there is no reason to change the grammar. If not, the learner will select a parameter at random in order to switch it. In contrast to the original approach it is not necessary that the sentence is generated by the new grammar. Then, local maxima cannot exist, because if it is not necessary to analyze input sentences in the prospective grammar, then the learner is legitimated to move away from a local maximum to any other neighboring grammar. Nevertheless, this also means that the learner can move away from the target grammar over and over again. This also leads to a psycholinguistically implausible algorithm.

The entries of the transition matrix $T_{\mathcal{M}_{nongr},5}$ for a non-greedy algorithm $\mathcal{M}_{nongr}$ are calculated as follows:

1. The probability for the learner to be triggered out of a hypothesis grammar G_h only dependends on the number of sentences in the target language, which are not analyzable by the hypothesis language. Thus, the diagonal entries—the properties to stay in a grammar—are calculated as $T_{\mathcal{M}_{nongr},5}[h,h] = 1 - \frac{|\mathcal{L}(G_T)\setminus\mathcal{L}(G_h)|}{|\mathcal{L}(G_T)|}$.

2. If the learner receives a triggering input, he choses a parameter at random (because of the single value constraint) and changes its value ($0 \mapsto 1$, $1 \mapsto 0$). Note, that for an example sentence it is not necessary to be in the prospective hypothesis language. Then, $T_{\mathcal{M}_{nongr},5}[h,h'] = \frac{1}{3} \cdot T_{\mathcal{M}_{nongr},5}[h,h]$ for all neighboring $G_{h'}$.

Hence, the transition matrix is

(62)
$$T_{\mathcal{M}_{nongr},5} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{2} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{1}{6} & \frac{1}{2} & 0 & \frac{1}{6} \\ 0 & 0 & \frac{5}{18} & 0 & \frac{5}{18} & 0 & \frac{1}{6} & \frac{5}{18} \\ 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Again, all local maxima disappear. The acquisition probability curves generated by the matrix above are shown in Figure 7. NB (cf. [NB 1996], [Niyogi 2006])

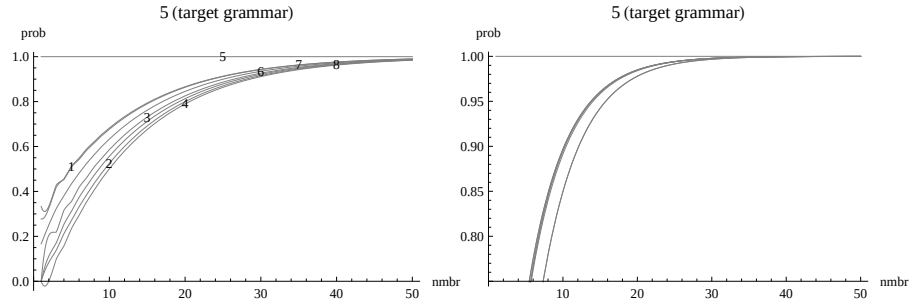


Figure 7: On the left: *Acquisition probabilities for target grammar 5 without the single value constraint. For all starting points the convergence behavior is similar and the learner is close to the target after approximately 30 example sentences. This is much faster than the single value learner.* On the right: *Acquisition probabilities for G_5 using a non-greedy algorithm. Also in this case there is no significant difference in the behavior of the curves. Note, that non-greedy learning is much slower than learning without single value constraint.*

also provided an analysis of non-greedy algorithms (random walks, random step algorithms) without the single value constraint, which appear to be the fastest algorithms concerning convergence to the target grammar. Hence, they argue against a TLA with constraints as above:

> *"Since the RSA* [random step algorithm] *is also superior in that it does not suffer from the same local maxima problem as TLA, the computational support for the TLA is by no means clear. Of course, it may be that future work will yield empirical support for the TLA, in the sense of independent evidence that children do use this procedure (given by the pattern of their errors, etc.), but this evidence is currently lacking, as far as we know."* [NB 1996, 178]

In this section we have seen how local maxima can be avoided, if a slightly modified learning algorithm is used. The learning behavior was examined for algorithms without the single value constraint as well es for non-greedy algorithms. Although changing the algorithm would afford a solution for the local maxima problem, the algorithm would then lack psycholinguistical adequacy. Since such a tradeoff is unsatisfiying, other solutions have to be investigated.

4.1.2 Default settings and parameter orderings

As changing the triggering learning algorithm brings us away from a linguistically plausible model, as seen in the previous section, the second attempt to solve the local maxima problem is the assumption of a default setting for some or all parameters, such that it is impossible reach local maxima. GW refer to [Hyams 1986], where a default setting for the null-subject parameter is described. Since the null subject parameter allows optional empty subjects, if set to 1, it generates subset and superset languages. That is, every grammar with

Table 5: *Problematic grammars. See also* [NB 1994, 176].

target	p_{spec}	p_{comp}	p_{V2}	potential dead ends	dead ends	local maxima
G_1	1	1	0	*learnable*		
G_2	1	1	1	*learnable*		
G_3	1	0	0	G_5, G_7		G_6, G_8
G_4	1	0	1	*learnable*		
G_5	0	1	0	G_1, G_3	G_4	G_2
G_6	0	1	1	*learnable*		
G_7	0	0	0	G_1, G_3		G_2, G_4
G_8	0	0	1	*learnable*		

$p_{NS} = 1$ can generate all sentences of the same grammar with the only difference that $p_{NS} = 0$. This consequently leads to local maxima, since no trigger exists for the latter grammar.

To solve this problem, the so-called *subset principle* is assumed, that is, *"given a choice between two parameter values, one of which generates a language which is a subset of the other, the child will initially assume the value which generates the subset language"* [Hyams 1986, 154]. GW adopt the idea of initial parameter settings in order to give a solution for the problem of local maxima.

As already stated in the introduction of part 4, GW rely on false assumptions, since there are more problematic grammars in the 3-parameter space than given in [GW 1994, 426, table 4]. A complete list of target grammars and associated problematic grammars determined using the methods developed by Niyogi and outlined in the previous part is given in table 5.

GW now differentiate between three possibilities [GW 1994, 4.1]:

1. *Default values for all parameters:* If for every parameter in the set there is an initial value, it turns out that in the 3-parameter space the learner always has the possibility to get to a local maximum. This means that there is no such default setting for all parameters, such that local maxima can be avoided.
   This follows immediately from the listing in table 5: Every grammar appears as problematic point for some target grammar. Grammars 1-4 are problematic for G_5 and G_7, grammars 5-8 are problematic for target grammar G_3. Hence, this solution is inappropriate.

2. *Unset initial value for all parameters:* In order to show that this solution is inappropriate GW argue as follows: At the beginning no parameters are set. Now, since the greediness constraint only allows to set one parameter at a time and because of the fact that for each grammar generated by the three parameters at least two parameter values are necessary to analyze an input sentence. Hence, the learner will never be able to analyze any sentences. So, at least one parameter must have a default value and this appriach must be abandoned.

3. *Unset initial values for some parameters and default values for other parameters:* The consequence of 1. and 2. now is that one or two parameters

are set (but not none or three). GW argue that there exists an initial parameter setting, which allows the learner to converge to all targets, based on the assumption that the only problematic points are local maxima and dead ends. Thus, they propose $p_{V2} = 0$ as initial parameter value.
If, however, also the remaining problematic points are considered, we run into problems: G_1, G_3 and G_5, G_7 are problematic for target G_5, G_7 and G_3, respectively. So, a second parameter has to be preset. For target G_5 or G_7 we had to set $p_{spec} = 0$ in order to exclude G_3. But, for target G_3 the spec parameter has to be set to 1, to ensure that the learner does not reach G_5 or G_7. Presetting the comp parameter instead would not help us out of our dilemma. As can be seen, initial values alone do not solve the local maxima problem.

Motivated by the problem of noise, i.e., data that cannot be analyzed by any possible grammar, GW argue for a more stable solution than only initial parameter settings. In fact, as shown in 3., initial parameter settings are not only insufficient in the presence of noise, but also under simple unnoisy learning conditions. To handle noisy data GW propose a more stringent way of setting parameters: GW suppose that there exists an ordering for the parameters, which ensures that certain parameters can only be set before or after other parameters. Thus, there is a period where only a subset of all parameters can be set.

Since GW come from initial parameter settings, this approach consists of three parts (cf. [GW 1994, 432f]):

1. The initial parameter setting is $p_{V2} = 0$, and the remaining parameters are set arbitrarily.

2. First, values for p_{spec} and p_{comp} are determined by the learner, while p_{V2} remains fixed.

3. *"After enough data have been processed so that the values for the base word order parameters can be set with confidence for those grammars without V2 movement, the learner then considers the alternative value for the V2 parameter"* [GW 1994, 433].

The second point, however, poses a problem: How should the learner know, when certain parameters are *"set with confidence"*? Since the learner cannot determine, whether he sets a parameter correctly[13], we have to assume a certain period, that is, a certain number of example sentences in our setting, where some parameters are fixed. Later on these parameters can be set as usual.

Thus, to model a parameter ordered learning process for, say, target grammar G_5, three parts are needed:

1. The initial parameter setting has to be modeled. Since we assume $p_{V2} = 0$, for an equally distributed setting of the remaining two parameters we obtain $\Pi_0 = (\frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}, 0)$. This means, that as starting points only G_1, G_3, G_5 and G_7 come into consideration.

[13] If he could, he would be better off determining the correct parameter setting immediately sparing a long-winding learning algorithm.

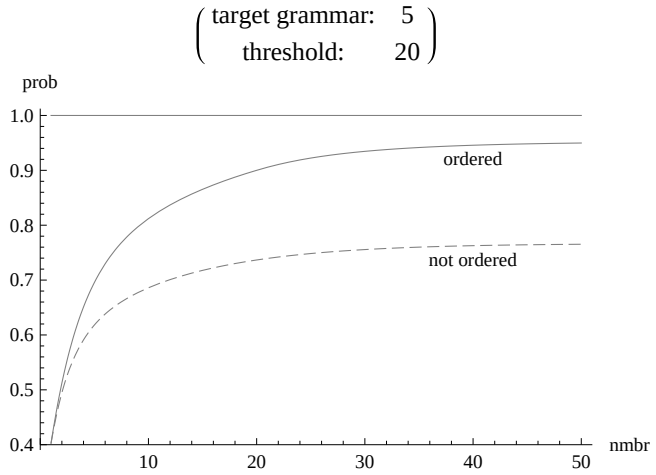


Figure 8: *Parameter ordered learning process*: *The dashed curve represents the learning behavior if no ordering but only the initial parameter setting* Π_0 *were used. The curve for the ordered learning process,* $\tau = 20$*, does not converge to* 1*. There is still a certain chance to get stuck in a local maximum, but the probability to do so is significantly decreased. It is clear that for smaller thresholds* τ*, this probability increases. For* $\tau = 0$*, in particular, both curves coincide.*

2. The value for p_{V2} is fixed, thus there will be no positive weighted outgoing arcs to any V2-grammar in the Markov chain. This, however, has no influence on the probability to go to another neighboring grammar. So, the probability to stay in the current grammar G_h increases exactly by the probability to go to the grammar $G_{h'}$, which emerges from the former changing the V2 parameter to 1.
   If, for example, $T_{\mathcal{M},5}[1,1] = \frac{1}{2}$ and $T_{\mathcal{M},5}[1,2] = \frac{1}{6}$ in the normal transition matrix, then, as $G_1 = (110)$ and $G_2 = (111)$, the corresponding entries in the transition matrix with a fixed value for the V2 parameter $T_{\mathcal{M}',5}$ are $\frac{4}{6}$ and 0, respectively.
   Let the specified period of time be τ. That is, at a threshold of τ example sentences the learner releases the fixed parameter value. Then at this threshold the probability distribution on the space of grammars is given as $\Pi_0 T^{\tau}_{\mathcal{M}',5}$.

3. Now, since V2 is not fixed any more, the learner can move though the space of grammars as usual, which is modeled by the Markov chain with transition matrix $T_{\mathcal{M},5}$. The associated probability distribution is $\Pi_0 T^{\tau}_{\mathcal{M}',5}$, thus the probability distribution after n example sentences, where $n > \tau$, is $\Pi_0 T^{\tau}_{\mathcal{M}',5} T^{n-\tau}_{\mathcal{M},5}$.

Figure 8 shows the learning behavior for $\tau = 20$. As the graph shows, there is yet a certain chance to reach a local maximum (after τ sentences) rather than the target language. Thus, we see that parameter ordered learning does

not guarantee learnability, but it provides an effective strategy to increase the probability to acquire the target language.

Though, there could be target grammars for which there exists no effective parameter ordering, or, as GW put it, *"*[w]*hether ordering solutions are available for all linguistically possible parameter spaces is an important open question"* [GW 1994, 438].

In this section, I have shown how initial parameter settings and parameter ordered learning processes can be modeled in terms of Markov chains. Moreover, I pointed out that despite the help of parameter orderings, there is still a positive probability not to converge to the target grammar, as the number of examples goes to infinity. That this might not be a big disadvantage is topic of the following section.

4.1.3 Unlearnable grammars - An excursion to diachronic syntax

One implication of the existence of local maxima could be that there are in fact unlearnable grammars. GW argue that in the 3-parameter setting local maxima are found also for grammars of existing languages, which are not unlearnable. Hence, they abort this solution.[14]

Moreover, in the previous section we have seen, that not even parameter orderings provide a solution for the local maxima problem. Nevertheless, one has to keep in mind that the process of language acquisition is finite, that is, only a finite number of example sentences can be given to the learner. In fact this means, that there will always be a positive probability not to be in the target grammar when the process of language acquisition ends. Besides, noisy data or data from a deviant language could hinder language acquisition and—in the worst case—trigger the learner out of the target grammar again.

At this point, evolutionary theories (such as [Nowak 2001], [Niyogi 2006], [KNM 2001], or [Yang 2000]) are tying in with the possibility to reach another grammar, different from the target. As a matter of fact, diachronic linguistic change over generations is possible only then, if parameters are set differently, that is, if the learning process is not successful.[15] Consider figure 9, which shows the probabilities not to acquire the target language but problematic grammars, with and without parameter orderings.

The question arises, if there is a relation between the existence of local maxima and the change of parameters over generations. Let us examine the just presented case of target grammar $G_5 = (010)$, which constitutes a fragment of English syntax. In particular, we find that G_5 and all other unlearnable grammars have value $p_{V2} = 0$, and all local maxima have value $p_{V2} = 1$. Hence, it is tempting to assume that is more likely to change from a non-V2 language to a V2 language over generations, since a learner exposed to a non-V2 language is more likely to get stuck in a V2 local maximum, rather than in another language (except, of course, the target).

But, however, observed parameter changes in diachronic syntax rather show the opposite: Take, for instance, the loss of V2 from Old French to Mod-

[14] Strictly speaking, this reasoning is not adequate for the simple fact that GW only argue within a rather small parameter space. It would be possible that enlarging the number of parameters some local maxima could disappear, inasmuch as the additional parameters can be used to escape the local maxima.

[15] Originally, this was motivated by the presence of noisy data rather than local maxima. NB introduced the idea of implementing noisy data in [NB 1996, ch. 4].

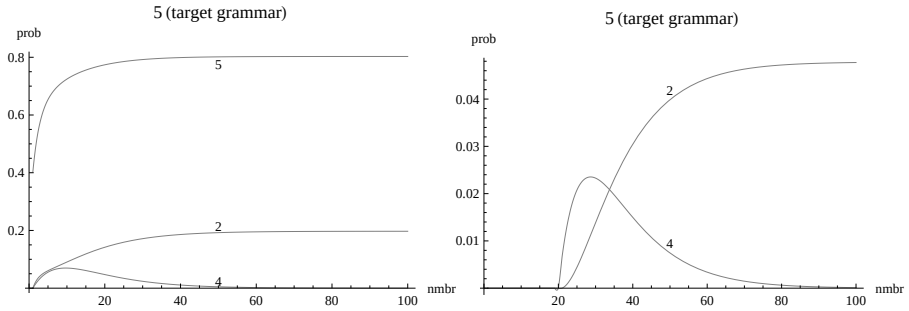


Figure 9: *Learning behavior of some problematic grammars.* On the left: *Learning without parameter ordering. The learner converges to the target G_5 with probability ~ 0.8, and to the local maximum G2 with probability ~ 0.2 after 50 example sentences.* On the right: *Learning guided by parameter ordering. The probability to reach the local maximum is significantly decreased (~ 0.048 after 100 example sentences). The curve for G_5 is not plotted here.*

ern French ((011) $\mapsto$ (010)), computationally analyzed in [Yang 2000], from Old English to Middle English to Modern English ((001) $\mapsto$ (011) $\mapsto$ (010); [Roberts 2007]), or *"also in most, if not all, other Romance languages at various points in the Medieval period"* [Roberts 2007, 108].

In addition, local maxima almost always differ from the target grammar at least two parameter values; this would mean that in an evolutionary process sequenced grammars differ significantly from one another. This, however, contradicts actual investigations: *"In many cases of language change, one finds that there are two variants (dialects, grammars) differing by a significant linguistic parameter that coexist in a population in varying proportions at different points in time"* [Niyogi 2006, 188].

Whether local maxima and language change correlate in some manner still remains an open question. Nevertheless, we have seen, that in an evolutionary context the criterion of learnability is somehow weakened: It is not of importance, whether a language can be acquired with probability 1 in the limit, rather mean convergence times are cruical for language change. I showed how parameter orderings have significant influence on the convergence time as well as on probability distributions.

On the other hand, the proposed solutions for the problem of local maxima remain unsatisfying: They are either psycholinguistically implausible or reduce the problem rather than completely solving it. Hence, a solution which eliminates local maxima entirely is needed.

4.1.4 No local maxima

Another possibility to avoid the local maxima problem is to assume that in UG there exists something yet unsufficiently investigated eliminating local maxima. GW offer two different approaches:

First, GW question the parameters used in their investigations, namely spec,

comp, and V2. That is, UG eventually consists of parameters other than the ones examined so far. But, *"the parameters that we have assumed here are fairly widely accepted, although with many variations. So until a better analysis of the range of data to be handled by these parameters is proposed, we will assume that these parameters are correct"* [GW 1994, 438].

Moreover, subset generating parameters like the null subject parameter can hardly be evaded using other parameters: The local maxima resulting from such parameters are due to clear subset properties rather than missing links between two or more grammars in the grammatical search space. Remind that the computation of the transition probabilies in the Markov model are based solely on set extensional critera.

Second, they suggest that there exists not yet analyzed data, which makes the learner escape the local maxima caused by the V2 parameter. GW give several proposals:

1. Degree-1 or higher order data (i.e., embedded sentences),
2. imperatives,
3. stranding effects, scrambling in non-V2 positions, V1 orders, fronting and topicalization,
4. sentence patterns including separable prefix verbs, and
5. intonation.

According to GW, solutions 1 and 2 are inappropriate, since both do not change the local maxima. In particular, to succsessfully analyze degree-1 or higher data, the learner has to correctly handle degree-0 data. It is exceedingly questionable, if degree-1 data can provide help to acquire the parameters values for less complex data.

The solutions in 3 require additional parameters, thus, it is likely that new local maxima appear (even if the original problematic states disappear). Solutions 4 and 5, instead, are *"data all of which include new kinds of information, so that no orders of any of the constituent patterns to be added are in the current parameter space"* [GW 1994, 440]. However, separable prefix verbs are not common to all languages, leaving intonation as proposal: GW argue that *"topicalized nonsubject pronouns must be stressed in +V2 languages"* [GW 1994, 439].

One has to keep in mind, that these solutions are proposed to deal with the local maxima caused by the V2 parameter and thus are ad hoc solutions, in a way. In order to handle the null subject problem, for example, other solutions have to be found.[16] The main problem of solutions 1 to 4 is that all are inherent in the syntactical system, and, as such, influence set extensions. Intonation, however, is the only approach implementing mechanisms belonging to another field of grammatical theory. To integrate such data in the formulation of the TLA is a detailed matter, since every local maxima generating parameter has to be accounted for. As this is goes beyond the scope of this thesis and for the matter that general solutions are more desirable than specific ones, I will omit

[16] It appears that the existence of null subject is dependent on θ-role assignment and finiteness, which is not captured under the 3-parameter fragment of UG. See [Hyams 1986, ch. 3.1.1] for an extensive discussion.

the modeling of intonation.

In this chapter I gave an analysis of several possible solutions for the problem of local maxima originally stated by GW. In particular for the most proposals I presented a mathematical model using the Markov chain approach developed in the preceding chapter.

In the case of algorithm changes, NB pointed out that the omission of one of the implemented constraints solves the problem. I described the changed algorithms in detail and illustrated the approach with convergence curves. The second approach, namely the integration of default settings and parameter orderings, was skipped by NB for the reason that it does not completely avoid the problem. However, I showed how parameter ordered algorithms in combination with appropriate default settings for parameter values can be modeled and demonstrated that the probability to get stuck in a local maximum is significantly decreased.

Furthermore, I posed the question, whether there is a relation between the existence of local maxima and diachronic language change. This question can be answered satisfactorily, only if some more diachronic data is investigated. At last I summarized, which of the remaining proposals of GW are reasonable, concluding that they are either inappropriate or not applicable in general. Coming from this, I go back to a syntactically not inherent solution, already brought in chapter 2.1, that is, the implementation of negative evidence.[17]

4.2 Negative triggers

The issue of negative evidence in child directed speech is a disputed one. Advocates of nativism often claim that negative evidence is not available to the child, since, first, child directed speech differs interculturally, and, second, for the reason of defending innate linguistic knowledge. Defenders of negative evidence, in contrast, refer to corrections and corrective repetitions in child directed speech. In [BS 1988] the existence of corrective feedback even is seen as an argument against nativism. This controversy mostly is caused by a misinterpretation of Gold's theorem, which I handled in 2.1.3 and 2.1.4, namely that negative evidence is formally equivalent to information presentation by informant.

In this chapter I will use negative evidence to solve the problem of local maxima, given in 4.1.3. More precisely, I will account for the following statement: If negative evidence is available to the child, such that the child can interprete input sentences labeled as negative evidence as trigger for a grammar change, then local maxima disappear. What I do not account for is, whether interpretable negative evidence does exist in child directed speech or not, as this still is an undecided matter in psycholinguistics, nor do I claim that negative evidence can be equated with informant type information presentation in Gold's sense, such that innate linguistic knowledge becomes formally redundant.

In particular, negative evidence only is one possible solution for the local maxima problem. There may be other appropriate solutions as well, which complement or even substitute negative evidence. In fact, in order to defend the TLA approach facing cultures, where negative evidence is obviously absent,

[17] [GW 1994] mentioned negative evidence in a footnote, but considered it not to be weighty enough to be an appropriate solution.

one has to find solutions other than negative evidence. These solutions may be implicit (e.g., the assumption that languages in such cultures are formally learnable, anyway), as well as explicit (e.g., the assumption of other triggering features, such as intonation). In this respect, I stick with [Pinker 1981, 29] stating that *"it* [...] *seems unlikely that negative evidence would be* necessary *for language acquisition to take place."* But, it may well be a *sufficient* condition for acquiring a language using a learning algorithm like the TLA.

The first section deals with different types of evidence in child directed speech in general. At the end of 4.2.1 I formulate a working hypothesis, which summarizes assumptions about the quantity and quality of negative evidence, in order to formally describe negative evidence in section 4.2.2. Then, I introduce some new definitions and the concept of negative triggers. Finally, I show how to integrate negative evidence into NB's Markov learning model.

In the third section I investigate how the assumption of different quantities affects learning times, and local maxima in particular. The final section, then, treats the local maxima problem given by the null-subject parameter, which has been introduced backwards in 2.2.3. I will demonstrate that for target grammar G_5, i.e., a fragment of German syntax, the behavior of the TLA using negative evidence and linguistic observations about the acquisition of the parameters in 2.2.3 show significant parallels.

4.2.1 Evidence in child directed speech

In psycholinguistic literature, evidence in child directed speech, or in speech that is heard by children in general, is split into three classes: *positive evidence*, (direct) *negative evidence*, and *indirect negative evidence* (cf. [Valian 2009], [Pinker 1981]).

1. *Positive evidence* is meant to be input sentences, a child hears during language acquisition. Pinker assumes that *"not all sentences heard by the child, nor all the parts of a sentence, will be used as input to his or her acquisition mechanisms. Presumably children encode most reliably the parts of sentences whose words they understand individually, and the whole sentences most of whose words they understand"* [Pinker 1981, 28]. This assumption is intuitive: A child would not use sentences as syntactically analyzable input, which are from another language it is usually exposed to.
   Note, however, that under this definition noisy data belongs to positive evidence, that is, sentences that are ungrammatical in the target language are treated as positive evidence by the child. This fact not excluded by the formulation of the TLA. As Valian states, input *"does contain a reasonable number of fragments and sentences without subjects about 5 per cent of the time"* [Valian 2009, 20]. An implementation of noise is crucial for modeling language change, as mentioned in 4.1.3.

2. *Negative evidence* subsumes *"parental corrections or subtle feedback when the child speaks ungrammatically"* [Pinker 1981, 28]. Negative evidence is direct in that it refers to utterances of the child, which implies (i) that negative evidence only occurs if the child is able to form sentences, their interlocutor can answer to, and (ii) that a child can identify negative evidence as such.

[BS 1988] divide parental responses in four classes, all of which the most relevant are *recasts*, i.e., corrective repetitions preserving the original meaning, and *expanded repetitions*, where major changes are arranged. Furthermore, they observed *adult clarification* questions, which contain no additional information.
[BMT 1995] show that it is questionable to count recasts as corrections, and further, to count corrections as negative evidence. Indeed, negative evidence in child directed speech is not without controversy. This will be discussed immediately.
It is important to note that negative evidence coincidentally serves as positive evidence: If a child hears a corrective utterance, this utterance is an input sentence, whether the child identifies it as correction or not.

3. *Indirect negative evidence*, in turn, *"is the absence of a structure that the child would expect to see, given a starting hypothesis"* [Valian 2009, 20]. This, however, implies that *"the child can infer the meaning of adults' utterances from their physical and discourse context and from meanings of the individual words in the sentence"* [Pinker 1981, 29]. Thus, indirect negative evidence is best implemented in a hypothesis testing learning algorithm.
Following Pinker, the support for indirect negative evidence mainly comes from the structure of child directed speach, namely, that it is simply structured (syntactically as well as semantically) and referring to the discourse context.
I will not account for indirect negative evidence in this thesis.

It is worthwhile noting that negative evidence must not be confused with learning by informant in Gold's framework. To see this it suffices to recognize that not every ungrammatical utterance also is corrected, and, what is more important, that the parents do not provide a complete list of labeled ungrammatical sentences, neither explicitly, nor implicitly. This is, because negative evidence only is a reaction to ungrammatical sentences uttered by the child. Thus, in order to guarantee information presentation in the Gold sense, a child would have to produce all possible sentences and wait for a labeling by the informant. This, clearly, is unrealistic.[18]

Furthermore, in child directed speech only a part of recasts and expanded repetitions follow ungrammatical sentences (see [Gordon 1990] and [BMT 1995], analyzing [BS 1988]):

(63) *"Child: He turned the light on.*

Parent: Yes, he turned on the light." [BMT 1995, 181]

The rest of negative feedback follows grammatical sentences, that is, negative evidence can be noisy as well. The child could infer that the correct utterance

[18] The existence or non-existence of negative evidence, then, has no impact on the assumption of innate linguistic knowledge in both ways: Neither does negative evidence form an obstacle for UG, nor would its presence predict that UG is unnecessary. I refer to Gordon's quotation on page 13.

I suppose that an informant-like information presentation rather is reached formally via hypothesis testing and indirect negative evidence: The child compares all possible hypothetic sentence structures with received input sentence structures to provide an appropriate labeling. However, then the problem of overgeneralization arises, since linguistic input hardly contains all possible structures (cf. [Valian 2009, 21]).

is ungrammatical due to an answer of type recast. Since noise in general is omitted in the analysis of the TLA in this thesis, I consequently will not handle noisy negative evidence.

Nevertheless, negative evidence is not neglected in literature. *"I think the assumption that negative evidence is not available to the child's learning mechanisms is warranted. There are, no doubt, cases in which parents correct their children"* [Pinker 1981, 29]. Indeed, Valian gives concrete percentage values for implicit negative evidence:

> *"Data from my laboratory, based on twenty-one child-mother pairs, suggest that parents provide 'implicit' corrections for about 25 per cent of children's ungrammatical utterances."* [Valian 2009, 21]

Note that of this 25 %, possibly only a part actually is identified as correction by the child. I will thus restrict further investigations to percentage values equal to or smaller than 25 %.

In summary, I will rely on the following assumptions in order to provide an implementation of negative evidence for the TLA:

(64) *Negative evidence: working hypothesis*

1. Negative evidence does exist in child directed speech to a certain extent.
2. Properties:
   (a) Negative evidence coincidentally also serves as positive evidence.
   (b) Children can identify corrective feedback as negative evidence to a certain extent.
   (c) Negative evidence appears only after a certain maturation time.
3. Restrictions:
   (a) There is no noisy negative evidence.
   (b) There is no indirect negative evidence.

4.2.2 Modeling negative evidence

In order to properly formalize negative evidence, such that all points in (64) are respected, I first introduce two new definitions:

(65) *Output sentences* and *answers*:

For a sequence of hypothesis grammars $(G_{h_1}, G_{h_2}, \ldots, G_{h_k})$, an *output sequence* is defined to be a sequence of sentences $(s_1, s_2, \ldots s_k)$, such that $s_j \in G_{h_j}$.

A sentence $s \in L_T$ is called an *answer* to a sentence s_j from the output sequence, if in the information sequence the index of s equals $j + 1$.

Up to now the TLA only relied on input sentences from an information string. Negative evidence, though, as outlined in the previous section is the reaction to an ungrammatical utterance of the child. I assume, that a child gives an output sentence after a potential change in linguistic behavior and befor hearing a

new input sentence. I will make this assumtion explicit in the reformulated algorithm.

I call the appearance of negative evidence a *negative trigger*, in order to avoid confusion. The concept will subsume implicit and explicit feedback, that is, there will be no differentiation within this framework between corrective repetitions on the one hand and explicit statements on grammaticality on the other.

As negative evidence is a property of certain sentences, which is *assumed* by hypothesis, there is no way to define negative triggers from within the framework, i.e., negative triggers cannot be defined formally via grammar or language relations, nor by correlations between input and output strings. A formal definition, which captures (64) entirely, would logically imply the existence of negative evidence. This, certainly, would be contradictory.

Thus, we only can formulate what it means for a learning procedure if a sentence has the propery of being a negative trigger:

(66) If a sentence s is a *negative trigger* for a grammar G_{h_j}, then the following holds:

1. $s \in L_T$,
2. in the learning sequence of hypothesis grammars $G_{h_j} \neq G_{h_{j+1}}$, where h_j is the index of the grammar where the learner resides when receiving s, and
3. s is an answer to a sentence $\sigma \in \mathcal{L}(G_{h_j})$, such that $\sigma \notin L_T$.

Let us consider what in fact the formualtion above does account for. The first point accounts for assumption (2a) in (64), since s is drawn from the target language. The second point means that the learner uses s to change his hypothesis, i.e., (2b). The third point accounts for (2c), since negative evidence only is a reaction to the child's utterances.

(3a) and (3b) are implicitly captured, since (64) provides no formulation for noisy negative evidence or indirect negative evidence. However, the formulation does not state that noisy negative evidence or indirect negative evidenc cannot exist. Moreover, it does not even account for the most crucial point, namely the existence of negative evidence. It only states how to use a sentence which is labeled as negative trigger.

Some more remarks: First, it is obvious that there exist no negative triggers for the target grammar by point (3). Second, if L_T is a subset of all other languages in the search space, then there exist no negative triggers, which is intuitively clear. Third, in contrast to global or local triggers, negative triggers are not related to parameter settings. That is, they only tell the learner whether a grammar is not the target, but they give no information about grammars that can generate the learner's utterance. This corresponds to adult clarification questions. Forth, the formulation makes no assumption about the relation between the incorrect utternace and the corrective sentence, that is, it is not said that the corrective sentence necessarily is the corrected form of the utterance. This is because within this framework there are no sentence-meaning relations. We are thus unable to make assertions about correct or uncorrect forms for a certain meaning. What we *could* state is, whether a sentence, say, "V O1 O2 S" with the constituents V, O1, O2, and S has one or more equivalent sentences

in another language, in that they also contain the same constituents. This is a refinement of the just to be stated algorithm, which I will not elaborate.

So, how to restate the algorithm? Let τ denote a time threshold, where the learner starts to generate sentences. This parameter ensures that negative triggers occur only if the child can produce complete sentences the interlocutor can react to.

(67) TLA with negative evidence

1. Start at some random point G_0 in the linguistic n-parameter space $\mathcal{G}^n$.
2. Provide a random output sentence $\sigma \in \mathcal{L}(G_h)$, if in the learning sequence of grammars for the index i of G_h holds that $i \geqslant \tau$.
3. Receive a random example sentence $s \in L_T$.
4. If $s \in \mathcal{L}(G_h)$, that is, if s is grammatical in the current hypothetical grammar, go back to step 2. Otherwise, continue.
5. Select a single parameter at random with probability $\frac{1}{n}$. Change its value ($0 \mapsto 1$, $1 \mapsto 0$), iff $s \in \mathcal{L}(G_{h'})$, where $G_{h'}$ denotes the altered hypothetical grammar, or if s is a negative trigger answering σ, where $h \neq T$. Go to step 2.

This formulation captures all effects of negative evidence given in (66). Note that still the learner only is allowed to make one step at a turn, but that in contrast to the original formulation, there are two possible reasons to change a parameter value: (positive) input sentences and negative triggers. Note furthermore that the two triggering events are connected by an *or* relation. That is, if one of the events triggers the learner out of a grammar, the other one is redundant. In particular, both events are not mutually exclusive. However, due to the *or* relation the learner is more likely to be triggered out of a sentence.

Then the probabilities for the transition matrix can be calculated as follows: Let be c the probability (i) that the interlocutor provides negative evidence if receiving an ungrammatical utterance and (ii) that the learner uses the information as negative trigger.

1. Calculate $\mathbb{P}(G_i \to G_j)$ for all i, j as in the original computation.
2. Calculate $q_i = \frac{|\mathcal{L}(G_i) \setminus L_T|}{|\mathcal{L}(G_i)|}$ for all i, i.e., the probability of providing an ungrammatical sentence if in G_i.
3. Compute the transition probabilities of the new transition matrix as $\mathbb{P}'(G_i \to G_j) = \mathbb{P}(G_i \to G_j) + \frac{q_i c}{n} \cdot \mathbb{P}(G_i \to G_i)$ for $i \neq j$, i.e., the original probability plus a probability to escape due to negative triggers, if the received sentence could not be used as trigger in the normal sense.[19]

[19] This can be seen by calculating

$$\begin{aligned}
&\mathbb{P}(\text{triggered normally AND triggered by negative trigger}) \\
&= \mathbb{P}(\text{triggered normally}) \cdot \mathbb{P}(\text{triggered by negative trigger}) \\
&= (1 - \mathbb{P}(\text{not triggered})) \cdot \mathbb{P}(\text{triggered by negative trigger}),
\end{aligned}$$

since both events are independent.

4. Compute the probabilities to stay in the hypothesis grammar for the new transition matrix as $\mathbb{P}'(G_i \to G_i) = (1 - q_i c) \cdot \mathbb{P}(G_i \to G_i)$.

I have to make two remarks on the calculations: First, for $i = T$, we find that $q_i = 0$, thus nothing can trigger the learner out of the target grammar. Second, which certainly is more important, it can be seen, that now local maxima are avoided, if only $c > 0$ and $q_i > 0$. The first statement holds, if there exists negative evidence interpreted by the learner. The second statement holds, if $L_T \nsubseteq \mathcal{L}(G_i)$, which is no restriction since if $L_T \subseteq \mathcal{L}(G_i)$, then there exist normal triggers, anyway.

In this section I showed how negative evidence can be formalized and how to integrate it in the TLA. I introduced negative triggers, which capture the properties of negative evidence. Further, I defined two parameters, τ and c, which denote a time threshold for sentence production and a measure for the availability and acceptability of negative evidence, respectively. Finally, I showed how to calculate a transition matrix for the TLA with negative evidence, and how negative triggers solve the problem of local maxima implicitly.

4.2.3 Convergence curves

In this section, I will demonstrate how learnability and convergence times vary dependent on the parameters c and τ, where the former describes the ratio of negative triggers accepted by the learner, and where the latter denotes the maturation time necessary for the child to produce sentences.

Let us start with parameter c. The convergence curves for target grammars G_5 and G_8, given in figure 10 were evaluated for different values for c, namely $c = 0$, $c = 0.1$, and $c = 0.25$. The first case corresponds to the absence of negative evidence, consequently the curves are the same as in 3.2. If c is increased, the local maxima disappear for target G_5: the curves for starting points G_2 and G_4 no more coincide with the x-axis. For $c = 0.1$ the learner reaches the target with high probability clearly after more than 50 example sentences. One can show, that it takes about 100 to 120 input sentences to acquire the target with sufficiently high probability. If, however, $c = 0.25$, then the curves converge significantly faster to 1 (about 60 example sentences).

In the case of target G_8 the variations are less significant: the convergence times vary between 40 and 50 example sentences for different c values. The comparison also shows that for $c = 0.25$, the differences between the curves for G_5 and G_8 are relatively small. Note, that the value of 25 % was mentioned by [Valian 2009].

Though c is constant in this analysis, this is not a given: the parameter may vary over time as a function of the number of example sentences. For example, one might consider c a decreasing function, in order to model the reduction of negative evidence in child directed speech. The adaption of child directed speech to the child's age is referred to as *fine-tuning* in literature (see for instance [Snow 1995]).

Varying parameter τ has a direct effect on convergence times: The parameter determines when negative evidence becomes available. Figure 11 shows that at time τ the learning behavior changes radically, which is more significant for large and less significant for small values. In particular, if a learner has reached a (former) local maximum ($p = 0.4$ in this case), τ directly influences the time for convergence to the target.

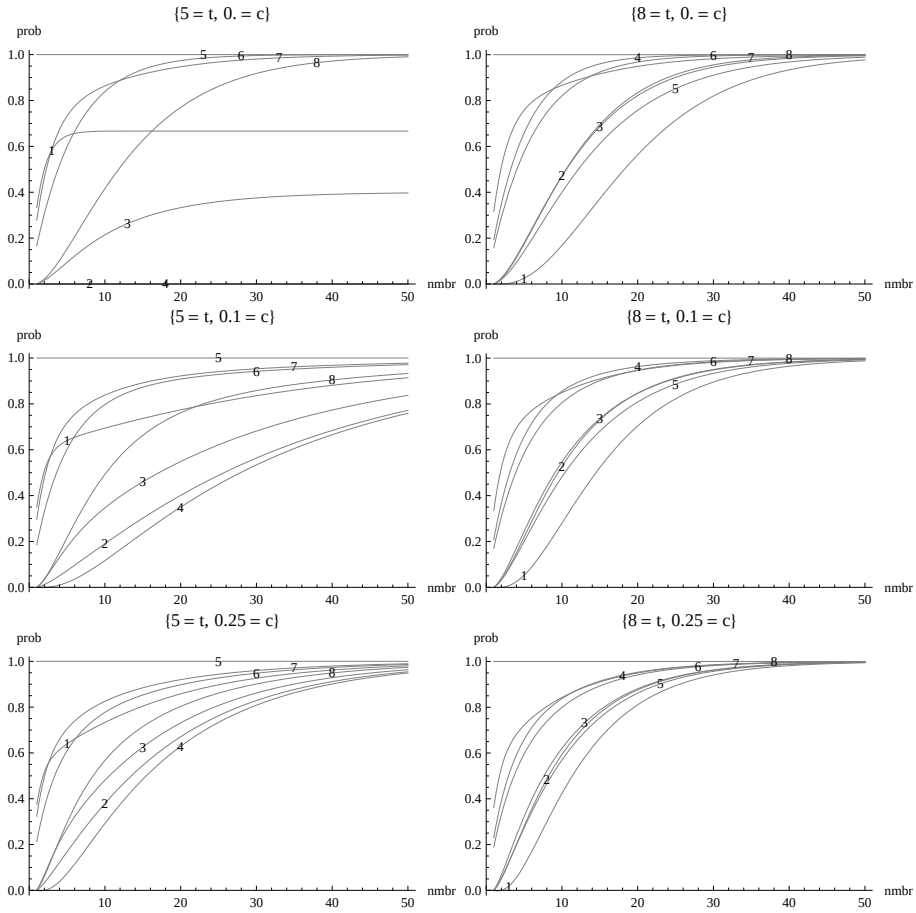


Figure 10: *Convergence curves for target grammars G_5 and G_8, where $\tau = 0$, for different values for c.*

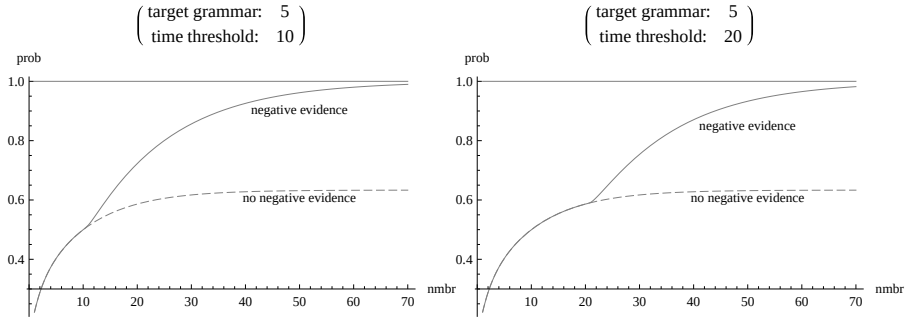


Figure 11: *Identification probabilities for target grammar G_5 for equally distributed starting grammars, where $\tau = 10$ (left) and $\tau = 20$ (right). In both cases, $c = 0.25$. The dashed curve shows the learning behavior without negative evidence ($c = 0$).*

Note that τ can also be interpreted as the point of incidence of negative evidence dependend on the informant rather than the maturation time for sentence production. Especially for this interpretation empirical investigations have to be made for different cultures.

Also parameter c poses empirical questions, I cannot account for in this section. But, if $c > 0$ is assumend, theoretical considerations about learnability may be made. In the last section I will show how negative triggers solve the subset problem in the TLA framework.

4.2.4 The subset problem: Setting the null-subject parameter

Subset languages are not only a problem for the TLA but also for the theory of language acquisition in general: If the target language L_T is a subset of a hypothetical language L_h, why should the learner, starting in the corresponding grammar G_h, change over to G_T during language acquisition? For the learner there is no reason to leave G_h, since all positive input sentences in the target language can also be generated by the hypothetical grammar. This problem is often referred to as *subset* or *superset problem* (cf. [Hyams 1986]).

In parametric frameworks the subset problem appears in particular for parameters allowing something optionally, such as the null-subject parameter that has been introduced in 2.2.3. In our restricted framework the null-subject parameter allows subjects to be empty. Note that in syntactic theory null-subjects are allowed only under certain conditions, which are not captured by the fragmental GB approach.

(68) *Null subject parameter*

S→Ø is a rule iff $p_{NS} = 1$

This rule is applied after all movements.

The linguistic space generated by *spec*, *comp*, *V2*, and the *null-subject* parameter is given in appendix A.2.

It is clear that the null-subject parameter generates problematic states and local maxima: If the target grammar has value $p_{NS} = 0$, then the same grammar with $p_{NS} = 1$ is a local maximum and all other null-subject grammars are problematic states, that is, they only lead to other problematic states or to local maxima. The transition matrix consists of two square submatrices, where one converges to the target (if learnable in the 3-parameter space), and where the other one converges to the local maximum, schematically described below. The remaining entries equal 0.

(69)
$$T_{\mathcal{M}} = \begin{pmatrix} [G_h \to G_T] & 0 \\ 0 & [G_h \to G_{loc\,max}] \end{pmatrix}$$

It should be remarked that due to the form of the transition matrix, there are no *potential dead ends* for a null-subject local maximum, i.e. a point in the space leading to the target and to the local maximum. For this reason, preset parameter values, as previously described, are a solution for local maxima generated by superset grammars. The so-called *subset principle*, that is, the assumption that a child never starts in a superset language generating grammar, accounts for this fact [Hyams 1986, 6.2.1].

Negative evidence, however, if assumend, also provides a solution for local maxima, since then for all null-subject grammars there is a positive probability to be triggered into the corresponding non-null-subject grammar. Figure 12 shows the convergence curves for grammars G_1 to G_{16} and target grammar G_8, a fragment of German. Clearly, G_9 to G_{16} are no problematic states any more: for all starting points the learner converges to the target grammar.

What is remarkable about figure 12 is the following: The curves for $i \neq T$ can be divided into three classes or bundles. Let G_4, G_6, G_7 form class A, G_1, G_2, G_3, G_5 form class B, and G_9 to G_{16} form class C. The curves in bundle A converge to 1 more rapidly that those in B, which in turn converge more rapidly than the curves in C. Class C clearly subsumes the null-subject grammars, class B consists mainly of non-V2 grammars (exception: G_2), while class A consists of V2 grammars (exception: neighbor grammar G_7).

Now consider the fact, that linguistic research discovered that children learning German first acquire the basic word order, while providing V2 errors until the age of 3 [Hyams 1986, 4.3], and null-subject errors until the age of 4 [Harmann 1996]. This means that at first, after setting *spec* and *comp*, the V2 parameter, and then finally the null subject parameter is set. [20]This ordering is completely analogous to the dynamics of the three classes in figure 12. It should be pointed out that under other approaches for the solution of the local maxima problem, we would not reveive such a result:

1. By removing the greediness constraint, all grammars would converge to 1 with about the same speed. The same holds for the single value constraint.

2. Using preset values for p_{NS}, we cannot observe the behavior of the learner starting in a null-subject grammar. In fact, there would be no explanation

[20]Note that this is an analysis only of convergence properties dependent on different starting points. Further investigations might capture probabilities to stay in grammars with certain mis-set parameters, mean convergence times and probability distributions thereof, as well as more realistic distributions on the set of input strings.

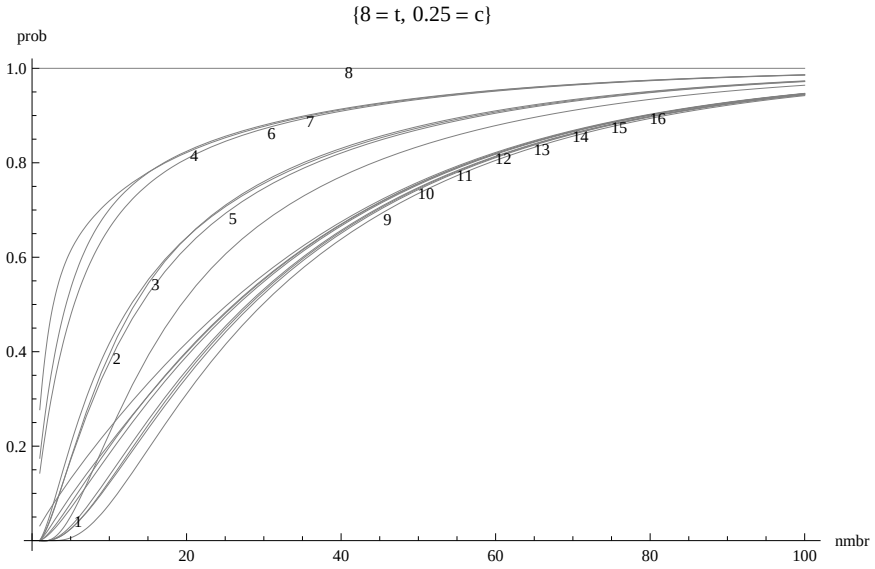


Figure 12: *Convergence curves for target grammar G_8 in a 4-parameter space, where $c = 0.25$ and $\tau = 0$.*

for making null-subject errors in early language; a phenomenon which also occurs in English.

3. All other solutions handled in the previous chapter either are ad hoc for the V2 parameter, or inappropriate.

Nonetheless, the assumption that the TLA combined with negative evidence is an adequate model for language acquisition in general is premature. The local maxima caused by the V2 parameter in the fragment of English still pose questions. But, the convergence of data from linguistic research and predictions made by the Markov formulation of the TLA in some respects makes the decision to abandon the considered computational approach premature as well.

5 Conclusion

Let us return to the initial task we were confronted with in the introductory words, namely the search for an appropriate theory of language acquisition, which can account for all empirical observations (remind figure 1). Clearly, the triggering learning algorithm is no such theory, rather it is a theory motivated model. The underlying theory is, as pointed out, nativism in the broader and GB combined with the existence of triggering sentences in the narrow sense. The criterion for rating the learning model is adequacy, thus predictions made by the TLA have to be examined and compared with empirical observations.

In the first chapter, two objections emerged in order to argue against Gold's straight forward language learning model (see 2.1). [Gordon 1990] pointed out that it lacks linguistical adequacy, since (i) a learner could guess a grammar by chance, and (ii) language identification time could be unrealisticly long. Indeed, Gold's ambition was not to make assertions about identifications times but about learnability in general—as originally it was the case for GW's TLA (see 3.1). Thus, to obtain a model whose predictions can be meaningfully compared with empirical data, and hence to account for Gordon's objections, help in form of NB's Markov chain formulation is needed (handled in 3.1.4).

However, success is short-lived, since the problem of local maxima arises. For [Yang 2004] this is one reason to prefer a variational learning approach to triggering. Certainly, the existence of local maxima is unsatisfying, but marginal changes might generally solve the problem; in short, they do not, as I elaborated in 4.1.

In a—not uncontroversal—way, the negative evidence approach suggests itself, since Gold's well-known theorem relates negative evidence-like information presentation with strong learnability abilities. Note, however, that the approach itself is not controversal because of Gold's theorem (I pointed out that negative evidence in child directed speech has nothing to do with formal information presentation via informant), but due to nescience about the effect of negative evidence in cognitive terms.

In contrast, from a purely mathematical point of view, modeling negative evidence is comfortable for two reasons: First, its implementation is straight forward, and second because of simply relying on some working hypotheses outlined in 4.2. Yet, these assumptions have to be proven.

When combining the TLA with negative evidence the results, nevertheless, are convincing, since computational predictions coincide with empirical data as demonstrated in the last section. So, what can be concluded from this observation? In the first instance it provides support for the TLA being an adequate language learning model. The answer for the question, whether the results support negative evidence, is ambiguous. They *do* support the existence of interpretable negative triggers, in that it can be shown that the results are yielded by a psycholinguistically motivated algorithm making use of negative evidence. They do *not* support their existence, as other kinds of triggers can lead to exactly the same results, if only they are formally equivalent. In fact, such triggers are necessary to account for language learning situations in cultures where negative feedback is obviously lacking. A proposal for equivalent triggering data may be some kind of indirect negative evidence, or observation of parental self-corrections.

Yet unconcerned are noisy evidence, positive as well as negative, and more

realistic probability distributions on the set of input sentences, in order to refine the algorithm. It is obvious that such an improvement is cruical for the computational investigation of diachronic language change, a much discussed topic in recent linguistic research.

The most important conclusion, anyhow, is the surprising support for the existence of problematic grammars the learner faces during language acquisition process. To be precisely, the results do not suggest that there are local maxima (clearly—they disappear in the modified algorithm), but that there are certain parameter settings the learner has—sooner or later solvable, but nevertheless apparent—problems to escape. That such problematic parameter settings exist is shown by psycholinguistic language acquisition studies. No other proposed solutions for the local maxima problem allow comparable observations.

That is, to put it into other words, local maxima are not a problem for the triggering learning algorithm, but rather yield a foundation for its psycholinguistical adequacy.

A Appendix: Mathematica codes

For all computations Mathematica 5.0 was used.

A.1 3-parameter languages and grammars

In the code, $G[1]$ to $G[8]$ are the parameter settings for a 3-parameter space (binary parameters). Below, $L[1]$ to $L[8]$ denote the corresponding languages, that is, the extensions for each of the grammars. In Mathematica subjects of the form "..." are seen as single list elements.

```
G[1] = {1, 1, 0};
G[2] = {1, 1, 1};
G[3] = {1, 0, 0};
G[4] = {1, 0, 1};
G[5] = {0, 1, 0};
G[6] = {0, 1, 1};
G[7] = {0, 0, 0};
G[8] = {0, 0, 1};

L[1] = {"V S", "V O S", "V O1 O2 S", "AUX V S", "AUX V O S",
        "AUX V O1 O2 S", "ADV V S", "ADV V O S", "ADV V O1 O2 S",
        "ADV AUX V S", "ADV AUX V O S", "ADV AUX V O1 O2 S"};
L[2] = {"S V", "S V O", "O V S", "S V O1 O2", "O1 V O2 S",
        "O2 V O1 S", "S AUX V", "S AUX V O", "O AUX V S",
        "S AUX V O1 O2", "O1 AUX V O2 S", "O2 AUX V O1 S", "ADV V S",
        "ADV V O S", "ADV V O1 O2 S", "ADV AUX V S", "ADV AUX V O S",
        "ADV AUX V O1 O2 S"};
L[3] = {"V S", "O V S", "O2 O1 V S", "V AUX S", "O V AUX S",
        "O2 O1 V AUX S", "ADV V S", "ADV O V S", "ADV O2 O1 V S",
        "ADV V AUX S", "ADV O V AUX S", "ADV O2 O1 V AUX S"};
L[4] = {"S V", "O V S", "S V O", "S V O2 O1", "O1 V O2 S",
        "O2 V O1 S", "S AUX V", "S AUX O V", "O AUX V S",
        "S AUX O2 O1 V", "O1 AUX O2 V S", "O2 AUX O1 V S", "ADV V S",
        "ADV V O S", "ADV V O2 O1 S", "ADV AUX V S", "ADV AUX O V S",
        "ADV AUX O2 O1 V S"};
L[5] = {"S V", "S V O", "S V O1 O2", "S AUX V", "S AUX V O",
        "S AUX V O1 O2", "ADV S V", "ADV S V O", "ADV S V O1 O2",
        "ADV S AUX V", "ADV S AUX V O", "ADV S AUX V O1 O2"};
L[6] = {"S V", "S V O", "O V S", "S V O1 O2", "O1 V S O2",
        "O2 V S O1", "S AUX V", "S AUX V O", "O AUX S V",
        "S AUX V O1 O2", "O1 AUX S V O2", "O2 AUX S V O1", "ADV V S",
        "ADV V S O", "ADV V S O1 O2", "ADV AUX S V", "ADV AUX S V O",
        "ADV AUX S V O1 O2"};
L[7] = {"S V", "S O V", "S O2 O1 V", "S V AUX", "S O2 O1 V AUX",
        "ADV S V", "ADV S O V", "ADV S O2 O1 V", "ADV S V AUX",
        "ADV S O V AUX", "ADV S O2 O1 V AUX"};
L[8] = {"S V", "S V O", "O V S", "S V O2 O1", "O1 V S O2",
        "O2 V S O1", "S AUX V", "S AUX O V", "O AUX S V",
        "O1 AUX  S O2 V", "O2 AUX S O1 V", "ADV V S", "ADV V S O",
        "ADV V S O2 O1", "ADV AUX S V", "ADV AUX S O V", "S AUX O2 O1 V",
        "S O V AUX", "ADV AUX S O2 O1 V"};
```

A.2 4-parameter languages and grammars

The lists are an extension of the lists in the previous section: The parameter settings are extended by one additional parameter, p_{NS}, resulting in 16 different grammars. The corresponding generated languages $L[1]$ to $L[8]$ are the same as in A.1, hence, they are not repeated here. This is, because the null subject

parameter generates superset languages for $L[1]$ to $L[8]$, noted as $L[9]$ to $L[16]$, respectively, in the code.

```
G[1] = {1, 1, 0, 0};
G[2] = {1, 1, 1, 0};
G[3] = {1, 0, 0, 0};
G[4] = {1, 0, 1, 0};
G[5] = {0, 1, 0, 0};
G[6] = {0, 1, 1, 0};
G[7] = {0, 0, 0, 0};
G[8] = {0, 0, 1, 0};
G[9] = {1, 1, 0, 1};
G[10] = {1, 1, 1, 1};
G[11] = {1, 0, 0, 1};
G[12] = {1, 0, 1, 1};
G[13] = {0, 1, 0, 1};
G[14] = {0, 1, 1, 1};
G[15] = {0, 0, 0, 1};
G[16] = {0, 0, 1, 1};

L[9] = {"V S", "V O S", "V O1 O2 S", "AUX V S", "AUX V O S",
        "AUX V O1 O2 S", "ADV V S", "ADV V O S", "ADV V O1 O2 S",
        "ADV AUX V S", "ADV AUX V O S", "ADV AUX V O1 O2 S",
        "V", "V O", "V O1 O2", "AUX V", "AUX V O", "AUX V O1 O2",
        "ADV V", "ADV V O", "ADV V O1 O2", "ADV AUX V", "ADV AUX V O",
        "ADV AUX V O1 O2"};
L[10] = {"S V", "S V O", "O V S", "S V O1 O2", "O1 V O2 S",
        "O2 V O1 S", "S AUX V", "S AUX V O", "O AUX V S",
        "S AUX V O1 O2", "O1 AUX V O2 S", "O2 AUX V O1 S", "ADV V S",
        "ADV V O S", "ADV V O1 O2 S", "ADV AUX V S", "ADV AUX V O S",
        "ADV AUX V O1 O2 S", "V", "V O", "O V", "V O1 O2", "O1 V O2",
        "O2 V O1", "AUX V",  "AUX V O", "O AUX V", "AUX V O1 O2",
        "O1 AUX V O2", "O2 AUX V O1", "ADV V", "ADV V O", "ADV V O1 O2",
        "ADV AUX V", "ADV AUX V O", "ADV AUX V O1 O2"};
L[11] = {"V S", "O V S", "O2 O1 V S", "V AUX S", "O V AUX S",
        "O2 O1 V AUX S", "ADV V S", "ADV O V S", "ADV O2 O1 V S",
        "ADV V AUX S", "ADV O V AUX S", "ADV O2 O1 V AUX S", "V",
        "O V", "O2 O1 V", "V AUX", "O V AUX", "O2 O1 V AUX", "ADV V",
        "ADV O V", "ADV O2 O1 V", "ADV V AUX", "ADV O V AUX",
        "ADV O2 O1 V AUX"};
L[12] = {"S V", "O V S", "S V O", "S V O2 O1", "O1 V O2 S",
        "O2 V O1 S", "S AUX V", "S AUX O V", "O AUX V S",
        "S AUX O2 O1 V", "O1 AUX O2 V S", "O2 AUX O1 V S",
        "ADV V S", "ADV V O S", "ADV V O2 O1 S", "ADV AUX V S",
        "ADV AUX O V S", "ADV AUX O2 O1 V S", "V", "O V", "V O",
        "V O2 O1", "O1 V O2", "O2 V O1", "AUX V", "AUX O V",
        "O AUX V", "AUX O2 O1 V", "O1 AUX O2 V", "O2 AUX O1 V", "ADV V",
        "ADV V O", "ADV V O2 O1", "ADV AUX V", "ADV AUX O V",
        "ADV AUX O2 O1 V"};
L[13] = {"S V", "S V O", "S V O1 O2", "S AUX V", "S AUX V O",
        "S AUX V O1 O2", "ADV S V", "ADV S V O", "ADV S V O1 O2",
        "ADV S AUX V", "ADV S AUX V O", "ADV S AUX V O1 O2", "V",
        "V O", "V O1 O2", "AUX V", "AUX V O", "AUX V O1 O2", "ADV V",
        "ADV V O", "ADV V O1 O2", "ADV AUX V", "ADV AUX V O",
        "ADV AUX V O1 O2"};
L[14] = {"S V", "S V O", "O V S", "S V O1 O2", "O1 V S O2",
        "O2 V S O1", "S AUX V", "S AUX V O", "O AUX S V", "S AUX V O1 O2",
        "O1 AUX S V O2", "O2 AUX S V O1", "ADV V S", "ADV V S O",
        "ADV V S O1 O2", "ADV AUX S V", "ADV AUX S V O",
        "ADV AUX S V O1 O2", "V", "V O", "O V", "V O1 O2", "O1 V O2",
        "O2 V O1", "AUX V", "AUX V O", "O AUX V", "AUX V O1 O2",
        "O1 AUX V O2", "O2 AUX V O1", "ADV V", "ADV V O", "ADV V O1 O2",
        "ADV AUX V", "ADV AUX V O", "ADV AUX V O1 O2"};
```

```
L[15] = {"S V", "S O V", "S O2 O1 V", "S V AUX", "S O2 O1 V AUX",
         "ADV S V", "ADV S O V", "ADV S O2 O1 V", "ADV S V AUX",
         "ADV S O V AUX", "ADV S O2 O1 V AUX", "V", "O V", "O2 O1 V",
         "V AUX", "O2 O1 V AUX", "ADV V", "ADV O V", "ADV O2 O1 V",
         "ADV V AUX", "ADV O V AUX", "ADV O2 O1 V AUX"};
L[16] = {"S V", "S V O", "O V S", "S V O2 O1", "O1 V S O2",
         "O2 V S O1", "S AUX V", "S AUX O V", "O AUX S V", "O1 AUX S O2 V",
         "O2 AUX S O1 V", "ADV V S", "ADV V S O", "ADV V S O2 O1",
         "ADV AUX S V", "ADV AUX S O V", "S AUX O2 O1 V", "S O V AUX",
         "ADV AUX S O2 O1 V", "V", "V O", "O V", "V O2 O1", "O1 V O2",
         "O2 V O1", "AUX V", "AUX O V", "O AUX V", "O1 AUX O2 V",
         "O2 AUX O1 V", "ADV V", "ADV V O", "ADV V O2 O1", "ADV AUX V",
         "ADV AUX O V", "AUX O2 O1 V", "O V AUX", "ADV AUX O2 O1 V"};
```

A.3 Transition matrix for TLA

The code represents the calculation of a transition matrix as described in 3.1.4. The value t for the target grammar, where $1 \leq t \leq 8$ or $1 \leq t \leq 16$, is not defined. Although a function for the Hamming distance is not difficult to define (count the number of differing positions) or implemented in more recent Mathematica versions, I used an alternative but in this case equivalent formulation via norms.

```
p = Length[G[t]];
n = 2^p;
m = Length[L[t]];

Do[
  L2[i] = Intersection[L[t], L[i]],
  {i, 1, n}
  ]

P = Table[0, {i, 1, n}, {j, 1, n}];

Do[
  If[Abs[Norm[G[i] - G[k], 1]] == 1,
    P[[i, k]] = 1/p*Length[Complement[L2[k], L2[i]]]/m
    ],
  {i, 1, n}, {k, 1, n}
  ]

Do[
  P[[i, i]] = 1 - Total[Table[P[[i, j]], {j, 1, n}]],
  {i, 1, n}
  ]
```

A.4 Transition matrix for TLA containing negative evidence

The code contains the computing algorithm as given in 4.2.2 in order to calculate a transition matrix for the TLA making use of negative evidence. The parameter $c \in [0, 1]$ is not defined and determines the relative frequency of sentences, valued as negative evidence, in PLD. As above, t indexes the target grammar.

```
p = Length[G[t]];
n = 2^p;
m = Length[L[t]];

Do[
```

```
  L2[i] = Intersection[L[t], L[i]],
  {i, 1, n}
  ]

P = Table[0, {i, 1, n}, {j, 1, n}];

Do[
  If[Abs[Norm[G[i] - G[k], 1]] == 1,
    P[[i, k]] = 1/p*Length[Complement[L2[k], L2[i]]]/m
    ],
  {i, 1, n}, {k, 1, n}
  ]

T = Table[0, {i, 1, n}, {j, 1, n}];

Do[
  Q[i] = Length[Complement[L[i], L[t]]]/Length[L[i]],
  {i, 1, n}
  ]

Do[
  T[[i, i]] = (1 - c*Q[i])*(1 - Total[Table[P[[i, j]], {j, 1, n}]]),
  {i, 1, n}
  ]

Do[
  If[Abs[Norm[G[i] - G[k], 1]] == 1,
    T[[i, k]] = P[[i, k]]
        + c*Q[i]/p*(1 - Total[Table[P[[i, j]], {j, 1, n}]])
    ],
  {i, 1, n}, {k, 1, n}
  ]
```

B Glossary

Chomsky hierarchy. When formalizing the notion of a language NOAM CHOMSKY proposed a hierarchy of types of →formal languages ordered by the generative capacity of the associated grammars. The different types are (i) →recursively enumerable languages (*type 0*), (ii) →context sensitive languages (*type 1*), (iii) →context free languages (*type 2*) and (iv) regular languages (*type 3*). In automata theory these types are related to (i) *Turing machines*, (ii) *linear-bounded automata*, (iii) *pushdown automata* and (iv) *finite state automata*, respectively. They are of (i) *undecidable*, (ii) *polynomial,* (iii) *exponetial* and (iv) *linear* complexity.

Context free grammar. A →formal grammar G is defined as *context free* if it is →context sensitive and all substitutions are of the form $n \to a$, where $n \in N$ and $a \in (\Sigma \cup N)^*$. A →formal language L is said to be context sensitive if $L = \mathcal{L}(G)$. An example for a context free language is $\left\{s^k t^k \mid k \in \mathbb{N}\right\}$ with the rules $n \to snt$ and $n \to st$.

Context sensitive grammar. A →formal grammar G is said to be *context free* if all rules are of the form $a_1 n a_2 \to a_1 \gamma a_2$, where $n \in N$, $a_1, a_2, \gamma \in (\Sigma \cup N)^*$ and either γ is not the empty string ε or $S \to \varepsilon$ is a rule and S is never on the right hand side of a rule. A →formal language L is said to be context sensitive if $L = \mathcal{L}(G)$. The language $\left\{r^k s^k t^k \mid k \in \mathbb{N}\right\}$ has a context sentitive grammar.

Formal languages. The theory of formal languages is a mathematical theory about the form of abstracted languages. For a finite set Σ a *formal language* L is defined as a subset of Σ^*, where $\Sigma^* = \{a_1 a_2 ... a_m \mid a_1, a_2, ..., a_m \in \Sigma, m \in \mathbb{N}\}$. The set Σ is called the *alphabet* of L and Σ^* is the set of all possible strings that can be formed by concatenation using the elements of Σ, also called *Kleene closure.*

Formal grammar. A *formal grammar* is a quadruple $G = \langle S, N, \Sigma, R\rangle$ such that $N, \Sigma \neq \emptyset$, $N \cap \Sigma = \emptyset$, $S \in N$ and R is a finite *set of rules* of the form $a_1 n a_2 \to a_3$, where $n \in N$ and $a_i \in (\Sigma \cup N)^*$ for $1 \leq i \leq 3$. Here S is the *start symbol*, N is the *set of nonterminal symbols* and Σ is the *terminal alphabet.* We say that G *accepts* a string σ if can be derived from S by the rules in R using N and Σ. The (formal) language generated by G is defined by $\mathcal{L}(G) := \{\sigma \in \Sigma^* \mid G \, accepts \, \sigma\}$.

Global trigger. A *global trigger* is a certain type of →triggering data, which ensures that a given parameter is set correctly irrespective of all other parameter values. It is formally defined as follows: "A *global trigger* for value v of parameter P_i, $P_i(v)$, is a sentence S from the target grammar L such that S is grammatical if and only if the value for P_i is v, no matter what the values for parameters other than P_i are." [GW 1994, 409].

Greediness constraint. The *greediness constraint* is a psycholinguistically motivated restriction in →learnability theoretic algorithms. The greediness constraint predicts that the learner only chooses reasonable hypothetical grammars insofar as he must be able to analyze an input sentence using the prospective parameter

setting.

Hamming distance. For two binary strings s_1, s_2 the *Hamming distance* $d_{hamm}(s_1, s_2)$, named after RICHARD HAMMING, is defined as the number of positions for which s_1 differs from s_2. Thus, it is a measure for the number of required substitutions to change one string into another. For example, for $s_1 = (0110)$ and $s_2 = (1011)$ we obtain $d_{hamm}(s_1, s_2) = 3$. The hamming distance is a distance in terms of mathematical topology.

Inductive inference. *Inductive inference* is a mathematical theory of making predictions about not yet observed events based on already observed ones. The theory partially relies on statistics and probability theory and is widely used in algorithmic →learnability theory and artificial intelligence.

Language identification in the limit. Also *learning in the limit.* Language identification in the limit is an inductive inference model for language acquisition, originally developed by E. M. GOLD. Based on presented input sentences from the target grammar the learner determines hypothetical grammars to possibly identify the target grammar if an infinite number of input sentences is given. The model includes two variants for data presentation: *text* and *informant*, where the first one represents only positive examples and the latter one in a sense also embraces →negative evidence.

Learnability theory. *Learnability theory* in the narrow sense is a subtheory of computational learning theory, which deals with learning algorithms for →formal languages. Computational learning theory addresses the analysis of machine learning.

Local trigger. A *local trigger* is a type of →triggering data weaker than →global triggers. It requires values for all parameters except one to determine the value of the remaining parameter. Local triggers are defined in the following way: "Given values for all parameters but one, parameter P_i, a *local trigger* for value v of parameter P_i, $P_i(v)$, is a sentence S from the target grammar L such that S is grammatical if and only if the value for P_i is v." [GW 1994, 409].

Markov chain. A *Markov chain* $\mathcal{M}$, called after ANDREY MARKOV, is a stochastic process with the following property: Consider a sequence of trials with the possible outcomes $E_1, E_2, ..., E_n$. Let p_{jk} be the conditional probability for a pair (E_j, E_k) and let a_k be the probability for E_k at the initial trial ($\{a_1, ..., a_n\}$ is assumed to be a →probability distribution). Then $\mathbf{P}(E_{j_0}, E_{j_1}, ..., E_{j_n}) = a_{j_0} p_{j_0 j_1} p_{j_1 j_2} ... p_{j_{n-1} j_n}$, that is, only the preceding trial is relevant for a certain state E_k. This is often referred to as *Markov property.* A Markov chain can be equivalently described as a *directed, weighted, probablistic graph* where $E_1, E_2, ..., E_n$ are associated with the nodes and where the weights on the edges are given by p_{jk} for (E_j, E_k).

Negative evidence. In the theory of language acquisition *negative evidence* is referred to as the information presented to the learner whether a sentence is ungrammatical. It is assumed that this information is offered indirectly via *recasts* (the learner's utterances are rectified by the informant preserving the original meaning) or *expanded repetitions* (like recasts but additional information is added).

Poverty of stimulus. The *poverty of stimulus argument* states that the linguistic stimuli presented to a child during the process of first language acquisition do not suffice to constitute the target grammar. This is due to the fact that grammatical structures are far too complex to be identified relying only on *positive evidence* and the rather small amount of (if existing at all) →*negative evidence* given to the child. In the *nativist approach* the poverty of stimulus argument is one of the main reasons for the assumption of →universal grammar.

Principles and parameters. *Principles and parameters* (p&p) is a framework for generative grammar closely related to →universal grammar and motivated by the theory of language acquisition. It is supposed that the innate *language faculty* within the human mind consists of finitely many fixed rules (principles) identical in every language and a finite set of variable parameters, such that any existing grammar can be represented by adopting the parameter values. During the process of language acquisition all parameters are set to finally constitute the target grammar. Common p&p theories are e.g. *Government and Binding* (GB), *Head-driven Phrase Structure Grammars* (HPSG), *Optimality Theory* (OT) and the *Minimalist Program* (MP).

Probability distribution. For a countable sample space Ω a function $\mathbb{P} : 2^{\Omega} \to [0,1]$, where 2^{Ω} denotes the powerset of Ω, is called a *probability distribution*, if (i) $\mathbf{P}(\Omega) = 1$ and (ii) for all pairwise disjunct subsets $A_i \subseteq \Omega$ $(i \in I)$ holds that $\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i)$.

Unrestricted grammars. Any →formal grammar is *unrestricted*. A →formal language L is *recursively enumerable* if $L = \mathcal{L}(G)$ for a formal grammar G.

Regular grammars. A →formal grammar G is called *regular* if it is →context free and all substitutions are of the form $n \to a$, where $n \in N$ and $a \in (\Sigma \cup \{\varepsilon\}) \times (N \cup \{\varepsilon\})$. A →formal language L is regular if $L = \mathcal{L}(G)$. For example $\left\{s^k t^l \mid k, l \in \mathbb{N}\right\}$ is a regular language.

Single value constraint. The *single value constraint* is a psycholinguistically motivated constraint in →learnability theoretic algorithms. To make a learner behave conservatively the Single Value Constraint allows two sequenced hypothetical grammars to differ only in at most one parameter value. It prevents the learner from changing his linguistic behavior too vastly.

Transition matrix. The entries of a *transition matrix* $T_{\mathcal{M}}$ for a →Markov chain $\mathcal{M}$ is given by the transition probabilities p_{ij} for all pairs of states (E_i, E_j).

$$T_{\mathcal{M}} = \begin{array}{c} E_1 \to \\ E_2 \to \\ \vdots \\ E_n \to \end{array} \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{n1} & \cdots & \cdots & p_{nn} \end{pmatrix}$$

Thus, $T_{\mathcal{M}}$ is always a square matrix and all rows $(p_{ij})_{1\leq j\leq n}$ form →probalility distributions on the sequence of states.

Triggering data. In →principles and parameters approaches *triggering data* for a certain parameter are taken to be sentences from the target language that cannot be analyzed by the learner using the current hypothetical

grammar. One differentiates between →*local* and →*global triggers*.

Universal grammar. In the nativistic theory *universal grammar* (UG) consists of the innate →parameters and principles, which are common to all speakers. UG treats grammatical aspects that are characteristical for all human languages rather than explaining only some of them.

C Nomenclature

Σ	alphabet
Σ^*	set of possible strings
ε	empty string
$\mathcal{L}(G)$	language generated by grammar G
G_T	target grammar
L_T	target language
$\mathcal{G}^n$	n-parameter grammar space
d_{hamm}	Hamming distance
p_k	value of parameter k, i.e., $p(k)$
p_{comp}	value of complement parameter
p_{spec}	value of specifier parameter
p_{V2}	value of verb second parameter
p_{NS}	value of null-subject parameter
$T_{\mathcal{M}}$	transition matrix for Markov chain $\mathcal{M}$
$T_{\mathcal{M},T}$	transition matrix for target grammar G_T
Π_0	initial probability distribution on $\mathcal{G}^n$
Π_k	probability distribution on $\mathcal{G}^n$ after k examples

References

[Bavin 2009] Bavin, E. L., ed. (2009), *The Cambridge Handbook of Child Language*, Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paolo (2009); New York.

[BMT 1995] Bonamo, K. M., J. L. Morgan and L. Travis (1995), *Negative Evidence on Negative Evidence*, Developmental Psychology 31, 2: 180-197.

[BS 1988] Bohannon, J. N., L. Stanowicz (1988), *The Issue of Negative Evidence: Adult Responses to Children's Language Errors*, Developmental Psychology 24, 5: 686-689.

[Briscoe 2002a] Briscoe, E. J. (2002), *Grammatical Acquisition and Linguistic Selection*, in ed. E. J. Briscoe (2002).

[Briscoe 2002b] Briscoe, E. J., ed. (2002), *Linguistic Evolution through Language Acquisition*, Cambridge University Press (2002); Cambridge.

[Bronstein 2005] Bronstein, I. N., K. A. Semendjajew, G. Musiol, H. Mühlig (2005), *Taschenbuch der Mathematik*, Verlag Harri Deutsch, Frankfurt (2005); Frankfurt.

[Chomsky 1957] Chomsky, N. (1957), *Syntactic Structures*, Mouton de Gruyter, Berlin New York (2002); Berlin.

[Feller 1950] Feller, W. (1950), *An Introduction to Probability Theory and Its Applications, Volume I*, John Wiley & Sons, Inc, New York, London, Sydney (1968); New York.

[Snow 1995] Snow, C. E. (1995), *Issues in the Study of Input: Finetuning, Universality, Individual and Developmental Differences, and Necessary Clauses*, in P. Fletcher, B. MacWhinney (eds.), The Handbook of child language, Blackwell Publishing, Malden, Oxford, Carlton (2004); Oxford.

[GW 1994] Gibson, E., K. Wexler (1994), *Triggers*, Linguistic Inquiry 25: 355-407.

[Gold 1967] Gold, E. M. (1967), *Language Identification in the Limit*, Information and Control, 10: 447-474.

[GM 1997] Gopnik, A., A. Meltzoff (1997), *Words, Thoughts, and Theories,* MIT Press, Cambridge, Massachusetts, London (1998); Massachusetts.

[Gordon 1990] Gordon, P. (1990), *Learnability and Feedback*, Developmental Psychology 26, 2: 217-220.

[Grimshaw 1981] Grimshaw, J. (1981), *Form, Function, and the Language Acquisition Device*, in eds. C. L. Baker and J. J. McCarthy, The Logical Problem of Language Acquisition, MIT Press, Cambridge, Massachusetts, London (1981); Massachusetts.

[Harmann 1996] Harmann, C. (1996), *Null Arguments in German Child Language,* Language Acquisition 5, 3: 155-208.

[Hausser 2000] Hausser, R. (2000), *Grundlagen der Computerlinguistik*, Springer, Berlin, Heidelberg, New York (2000); Heidelberg.

[Howard 1998] Howard, A. (1998), *Lineare Algebra*, Spectrum Akademischer Verlag, Heidelberg, Berlin (2004); Berlin.

[Hyams 1986] Hyams, N. (1986), *Language Acquisition and the Theory of Parameters (Studies in Theoretical Psycholinguistics)*, Kluwer, Dordrecht, Norwell (1986); Dortrecht.

[Johnson 2004] Johnson, K. (2004), *Gold's Theorem and Cognitive Science*, Philosophy of Science 71: 571-592.

[KNM 2001] Komarova, N., P. Niyogi, M. Nowak (2001), *Evolutionary Dynamics of Grammar Acquisition*, Journal of Theoretical Biology 209: 43-59.

[Kracht 2003] Kracht, M. (2003), *The Mathematics of Language*, Mouton de Gruyter, Berlin, New York (2003); Berlin.

[Lasnik 1981] Lasnik, H. (1981), *Learnability, Restrictiveness, and the Evaluation Metric*, in eds. C. L. Baker and J. J. McCarthy, The Logical Problem of Language Acquisition, MIT Press, Cambridge, Massachusetts, London (1981); Massachusetts.

[NB 1994] Niyogi, P., R. C. Berwick (1994), *A Markov Language Learning Model for Finite Parameter Spaces*, Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Las Cruces, New Mexico (1994): 171-180.

[NB 1996] Niyogi, P., R. C. Berwick (1996), *A Language Learning Model for Finite Parameter Spaces*, Cognition 61, 1-2: 161-193.

[Niyogi 2006] Niyogi, P. (2006), *The Computational Nature of Language Learning and Evolution*, MIT-Press, Cambridge, Massachusets (2006); London.

[Nowak 2001] Nowak, M. A. (2001), *From Quasispecies to Universal Grammar*, Zeitschrift für Physikalische Chemie 216: 5-20.

[Pinker 1981] Pinker, S. (1981), *Language Learnability and Language Development*, Havard University Press, Cambridge, Massachusetts, London (1981); Massachusetts.

[PS 2006] Pollum, G., B. Scholz (2006), *Irrational Nativist Exuberance*, in ed. R. Stainton, Contemporary Debates in Cognitive Science, Blackwell Publishing, Malden, Oxford, Victoria (2006); Oxford.

[Roberts 2007] Roberts, I. (2007), *Diachronic Syntax*, Oxford University Press, Oxford, New York (2006); New York.

[Valian 2009] Valian, V. (2009), *Innateness and Learnability*, in ed. E. L. Bavin (2009).

[Wexler 1981] Wexler, K. (1981), *Some Issues in the Theory of Learnability*, in eds. C. L. Baker and J. J. McCarthy, The Logical Problem of Language Acquisition, MIT Press, Cambridge, Massachusetts, London (1981); Massachusetts.

[Williams 1981] Williams, E. (1981), *A Readjustment in the Learnability Assumptions*, in eds. C. L. Baker and J. J. McCarthy, The Logical Problem of Language Acquisition, MIT Press, Cambridge, Massachusetts, London (1981); Massachusetts.

[Yang 2000] Yang, C. D. (2000), *Internal and external forces in language change*, Language Variation and Change 12: 231-250.

[Yang 2004] Yang, C. D. (2004), *Universal Grammar, statistics or both?*, Trends in Cognitive Science 8, 10: 451-456.

[Zuidema 2003] Zuidema, W. (2003), *How the poverty of the stimulus solves the poverty of the stimulus*, in eds. S. Becker et. al., Advances in Neural Information Processing Systems, 15, MIT Press, Cambridge (2003): 51-58.

Abstract

Language acquisition theory suggests a variety of formal models that provide a description of the search of a certain target grammar the learner is supposed to learn based on linguistic input given to the learner, which are adequate to a greater or lesser extent. These models vary in their underlying linguistic frameworks as well as in different underlying approaches to the problem of language acquisition as such.

This thesis focusses on *parameters and principles* theory (more precisely, *Government and Binding*, GB) as linguistic framework and *trigger*s, that is, input sentences that are ungrammatical in a learner's hypothetical grammar, as basic language acquisition concept. The modeling of the search for the target grammar was described by Gibson and Wexler in an algorithmic—and, in particular, formal—way, thus being referred to as *Triggering Learning Algorithm* (TLA). Once the TLA is confronted with language learning situations based on few elementary parameters of GB-theory, problems arise, as the learner can get stuck in certain grammars, denoted as *local maxima*.

The thesis is concerned with the motivation for the TLA, its formulation, several solutions for the problem of local maxima, and qualitative analyses thereof. In order to properly analyze the algorithm, three preparative parts are worked out: First, for clarification the fundamental model of formal language learning proposed by Gold and its relation to natural language acquisition are examined. Second, the thesis comprises an outline of the grammatical framework, a three-parameter subsystem of GB, serving as field of application for the TLA. Third, the mathematical framework of Markov chains, originally proposed by Niyogi and Berwick, is introduced in order to provide a statistical analysis of the behavior of the algorithm.

As the argumentation will show, the accounts for the problem of local maxima, which have been investigated so far, are deficient in that they lead to inadequate predictions. Therefore, an alternative approach is developed, namely the implementation of *negative evidence*, that is, sentences, which show the learner that the hypothesized grammar is faulty, together with a corresponding formal description. Investigations of the modified algorithm will reveal predictions that coincide with results of natural language acquisition research.

Although not the main purpose of this thesis, the relation between formal modelings and linguistic theories is emphasized persistently. Psycholinguistically adequate computational language learning models are of significant importance not only for language acquisition but also for fields such as artificial intelligence or evolutionary linguistics.

Kurzfassung

Aus der Spracherwerbstheorie gehen zahlreiche – nicht immer adäquate – formale Modelle für die Beschreibung des auf sprachlichem Input basierenden Erlernens von Grammatiken hervor, welche sich allerdings sowohl wegen ihrer jeweils zugrundeliegenden linguistischen Theorien als auch aufgrund verschiedener Herangehensweisen an den Erwerb natürlicher Sprache voneinander unterscheiden.

Die vorliegende Diplomarbeit beschränkt sich auf *parameters and principles* (genauer: *Government and Binding*, GB) als zugrundeliegende Theorie und *trigger*, also Sätze, die nicht durch eine bestimmte hypothetische aber vom Erwerbsziel verschiedenen Grammatik analysierbar sind, als zugrundeliegendes Spracherwerbskonzept. Die hier behandelte und ursprünglich von Gibson und Wexler eingeführte Spracherwerbsmodellierung, der *Triggering Learning Algorithm* (TLA), ist formal algorithmisch. Dabei treten allerdings Probleme auf, selbst wenn man den Algorithmus auf wenige Parameter aus der GB-Theorie anwendet: Der Lernende gerät in problematische Grammatiken, welche als *lokale Maxima* bezeichnet werden, die er nicht mehr verlassen kann.

Die Arbeit befasst sich sowohl mit Motivation und Formulierung des TLA als auch mit diversen Lösungsansätzen für das Problem der lokalen Maxima, sowie qualitativen Untersuchungen dieser Ansätze. Um eine gründliche Analyse zu ermöglichen, finden sich drei vorbereitende Abschnitte in der Arbeit: Erstens wird das zugrundeliegende Modell von Gold für den Erwerb formaler Sprachen beleuchtet, zweitens wird der grammatikalische Rahmen beschrieben, in welchem sich der Algorithmus bewegt – nämlich ein Teilsystem von GB bestehend aus drei Parametern. Drittens enthält die Arbeit einen Abschnitt, der die Theorie der Markoffketten zweckdienlich einführt. Ursprünglich wurde die Modellierung durch Markoffketten von Niyogi und Berwick für die statistische Analyse des TLA entwickelt.

Es wird gezeigt, dass die bis jetzt vorgeschlagenen Lösungsansätze für das Problem der lokalen Maxima mangelhaft sind, da sie zu falschen Voraussagen führen, weshalb ein weiterer Ansatz erarbeitet und formal beschrieben wird, der *negative Evidenz* mit einbezieht. Damit sind Inputsätze gemeint, welche dem Lernenden zeigen, dass die hypothetische Grammatik nicht mit der zu erreichenden übereinstimmt. Der so modifizierte Algorithmus liefert Ergebnisse, die mit empirischen Untersuchung-en zum kindlichen Erstspracherwerb übereinstimmen.

Obwohl kein zentrales Ziel der Arbeit, so wird, um in der Literatur häufig vorkommenden Verwirrungen vorzubeugen, auf das Verhältnis zwischen formalen Modellen und linguistischen Theorien ein besonderes Augenmerk gelegt. Neben der Spracherwerbsforschung ist die genaue Formulierung psycholinguistisch adäquater Spracherwerbsmodelle unter anderem auch für die Gebiete der KI-Forschung oder der Linguistischen Evolutionstheorie von großer Bedeutung.

Curriculum vitae

1985-07-30	born in Vienna
09/1992—06/1996	Primary school, Volksschule, Scheibenbergstraße 63, 1180 Vienna
09/1996—06/2004	Grammar school specializing in humanities, Bundesgymnasium XVIII, Klostergasse 25, 1180 Vienna
07.06.2004	School-leaving examination, passed with distinction
10/2004—09/2005	Community service, St John Ambulance, 1180 Vienna
Since 10/2005	Studies of Theoretical and Applied Linguistics, and Mathematics, Vienna University
2007-07-10	First diploma examination, Theoretical and Applied Linguistics, passed with distinction
2008-04-26	First diploma examination, Mathematics, passed with distinction
2009-07-16	Second diploma examination, Theoretical and Applied Linguistics, passed with distinction
07/2009	Beginning to work on the diploma thesis supervised by Prof. Martin Prinzhorn