Wissenschaftliche Arbeit
zu Erlangung des akademischen Grades
Magister Artium
(Magisterarbeit)

Identification of syntactic patterns in large corpora and aspects of structure in Chintang child-surrounding speech

vorgelegt von
Taras Zakharko (MN 9629812)

im Juli 2009

am Institut für Linguistik
der Philologischen Fakultät
der Universität Leipzig

Gutachter:

Prof. Dr. Balthasar Bickel
Institut für Linguistik
Universität Leipzig
Beethovenstraße 15

Dr. Sabine Stoll
Department of Linguistics
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig, Germany

CONTENTS

Contents i

List of Tables iii

Preface iv

Acknowledgements v

Introduction. Scope and structure 1

I Item-based patterns and language acquisition 3

1.1 Overview . . . 4

1.2 The innateness hypothesis and its criticism . . . 7

1.2.1 Nativism vs. empiricism . . . 7

1.2.2 Poverty of the stimulus argument . . . 7

1.2.2.1 Lack of positive evidence . . . 9

1.2.2.2 Lack of negative evidence . . . 10

1.2.3 Intermediate summary . . . 11

1.3 Data-driven learning and the acquisition of lexical categories . . . 13

1.3.1 Cues in the acquisition of lexical categories . . . 13

1.3.2 Distributional learning . . . 14

1.3.3 Possibilities of data-driven learning . . . 16

1.4 Frequent patterns in the language . . . 17

1.4.1 Frequent patterns in the child language . . . 17

1.4.2 Frequent patterns in the child-directed language . . . 18

1.5 Outlook . . . 20

II Identification of frequent patterns in corpora 21

2.1 Overview . . . 22

2.2 Patterns in formal languages . . . 24

2.2.1 Preliminaries . . . 24

2.2.2 Regular languages . . . 25

2.2.3 Regular expressions . . . 27

2.2.4 Regular expression generation . . . 28
2.3 Restricting the pattern language . . . 31
2.3.1 Interesting and uninteresting patterns . . . 31
2.4 Constraining the pattern derivation rules . . . 33
2.4.1 The restricted pattern language . . . 34
2.5 The framework and its implementation . . . 36
2.5.1 String encoding . . . 37
2.5.2 Pattern generation . . . 38
2.5.3 Pattern filtering . . . 39
2.5.3.1 Filtering of patters with flexible gaps . . . 40
2.5.4 Filtering of mutually ambiguous patterns . . . 40
2.5.5 Intermediate summary . . . 41

III A case study: child-surrounding speech in Chintang 43
3.1 Chintang and its people . . . 44
3.1.1 Chintang from a typological perspective . . . 44
3.2 The corpus . . . 46
3.3 Pattern identification . . . 48
3.3.1 Corpus preparation . . . 48
3.3.2 Pattern generation and comparative chunk evaluation . . . 50
3.4 Pattern analysis . . . 54
3.4.1 Distribution . . . 54
3.4.2 Pattern types . . . 54
3.4.3 Utterance-initial core patterns . . . 56
3.4.4 Utterance-final core patterns . . . 57
3.5 Summary and outlook . . . 58

4 Conclusions and outlook 61
4.1 Pattern identification: outlook . . . 61
4.2 The hidden structure in the language . . . 63

A Globally frequent pattern list 65

B Summary in German 71

Bibliography 72

LIST OF TABLES

3.1 Data chunks (distribution and size) 49
3.2 Frequent patterns per chunk (unfiltered) 51
3.3 Frequent patterns per chunk (filtered) 51
3.4 Number of globally frequent patterns within the data chunks and their distribution 54
3.5 Pattern classification and counts 55
3.6 Number of lexical items within the patterns 56
3.7 Utterance-initial frequent morphemes 56
3.8 Utterance-final frequent morphemes 57

A.1 Utterance-initial patterns 65
A.2 Utterance-final patterns 66
A.3 Utterance-medial patterns 67

PREFACE

I have to admit that writing this thesis was both a spontaneous and a daring undertaking. Spontaneous, as it was only in march 2009 that I have decided — for a number of independent reasons — that it was time to graduate. Daring, because I had to research a topic in linguistics I have never dealt with before, design a computer algorithm of which I was not sure how it should look like — and even if it could be designed in the first place, analyze a language and write a thesis — all that in under three months. Needles to say, the outcome could have been much worth.

Surprisingly, the work on the thesis — despite being extremely tiring and very irritating at parts — was one of the most fun things I have ever done. With bewilderment, I came to understand myself and my way of thinking better. This thesis is about patterns. And while writing about patterns in the language I also started noticing patterns in my own mind, as the knowledge and the experience gathered in the past years started to form some distinct and elusive, but nevertheless coherent entity. I suddenly remembered the reason why I began studying linguistics in the first place.

ACKNOWLEDGEMENTS

There are many people without whom writing this thesis would be a great deal more difficult and the end result more displeasing. To this people I owe my deepest gratitude. The complete list would take more pages then the thesis, but there are some who deserve special notion.

Of course, there are my supervisors, Sabine Stoll and Balthasar Bickel. First, as they were the ones who provided me with the idea of writing this thesis in the first place, without them I would have missed a unique and very enriching experience. Additional credit goes to them for the patience they showed in answering my often disorganized questions.

I am grateful to the Chintang and Puma Documentation Project (CPDP) and its team for their hard work on the Chintang language acquisition corpus which I use in my thesis. The project is funded by Volkswagenstiftung as part of the DoBeS project (Grant № II/79 092, 2004-2008 PI Balthasar Bickel).

I also owe to Tyko Dirksmeyer, Alexander Hoffman, Juliane Böttger and Sebastian Sauppe (in no particular order) for their invaluable comments and suggestions. A special thanks goes to Ronja Drewes who has shown considerable tenacity trying to untangle my English spelling.

This page would be incomplete without mentioning Donald Knuth, for two reasons. First, I would like to thank him for his creation, the LaTeX typesetting system, which made the process of writing this thesis so much easier and its printed looks so clear. Second, and even more important, his works in computer science and views on computer programming as an art have significantly contributed to my development as a designer of computer algorithms.

I also would like to thank Prof. Dr. Manfred Droste of the University of Leipzig for his consultations about the current state of research on formal languages.

Also, very special gratitude goes to my dear friends and colleagues Eva Zimmerman and Doreen Georgi with whom I have shared the office during most of the writing. Their humor and frequent reminders about the necessity of hot meals have made the writing of this thesis clearly more fun.

And finally, I want to thank my mother for raising me in such an independent way and for her indefatigable emotional support.

Introduction. Scope and structure

The question how children acquire the language is without any doubt a central problem in linguistics. An initial — and straightforward — observation is that children, independent of race, intellect or social position, are able to learn the language perfectly in relatively short time, without it being taught to them in a systematic way. This seems like quite a feat, given the vast complexity of the system that language is — and considering the fact that a large number of brilliant minds have struggled for over a century to understand how the language actually works, with no definite answer till now — this feat is even more remarkable.

I would like to start by clearly defining the scope of this study. Recent studies in language acquisition (Cameron-Faulkner et al., 2003; Stoll et al., 2009; Lieven et al., 1997; Tomasello, 2003, e.g.) have shown that item-based patterns which occur frequently in the speech play an important role in the process of learning a language. Thus, there is an interest in studying such patterns. However, no general methodology for identification of such patterns has been proposed so far. This is a shortcoming which I want to address. Therefore, there are two basic problems I discuss in this thesis: the methodological one (identification of patterns) and the theoretical one (which concerns a broad spectrum of theories on language acquisition and language learnability). These problems are clearly independent, but it does not stop them to interact in a number of ways. Thus, the discussion of the theoretical problem motivates the methodological problem, and subsequently the solution of the methodological problem affects the theoretical problem. However, as far as my thesis is concerned, the methodological problem is clearly the central one. The basic reason behind this is that this problem is by far "smaller" then the theoretical complex it accompanies. If I had the means to provide a complete and working account for language acquisition without having to deal with the methodology first, I would clearly do so. As it is obviously not the case I have to settle for

a problem I can solve, hoping that my solution will aid the subsequent research on the big question. In summary, the theoretical discussion in my thesis should be strictly regarded as secondary in respect to my main goal of providing a practical solution to the pattern identification problem.

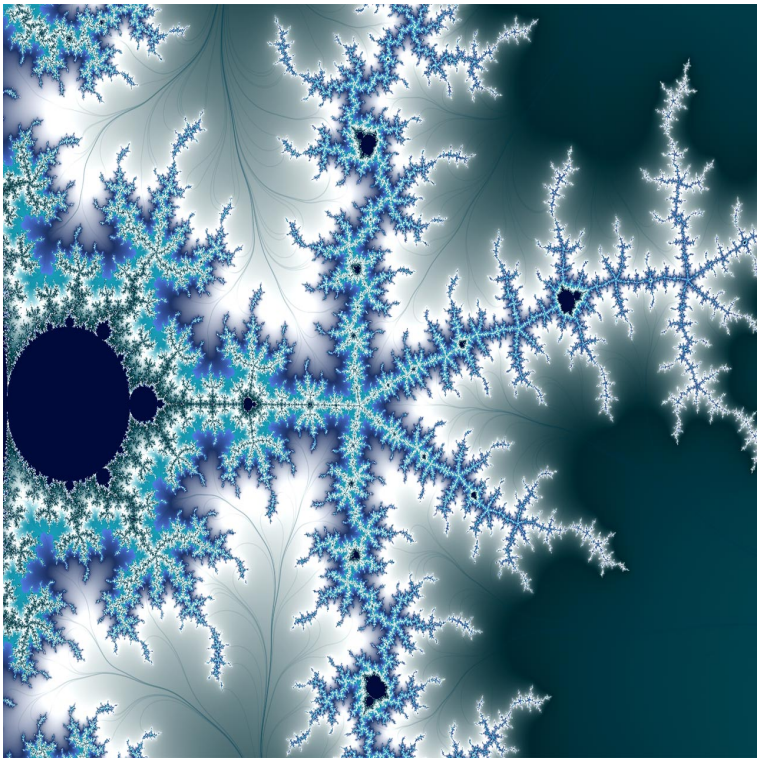
In the first chapter I lay out the theoretical background for my thesis, by discussing current problems and results in the language acquisition research. This brief overview, which by no mean is supposed to be complete, serves the purpose of outlining the significance in the frequent item-based patterns in the language and thus provides a starting point for the following discussion.

In the second chapter I discuss the problem of frequent pattern identification in language corpora. In many respects this is the core part of the thesis as it presents what I consider to be the most important result of my work. In short, by utilizing particular aspects of the formal language theory, I show how — under reasonable constrains — arbitrary patterns in a corpus can be identified and represented. A working practical framework for frequent pattern identification is presented, along with a detailed description of the details of its implementation.

In the final chapter, I take this framework for a test run by performing a case study. Here, I present an analysis of frequent morphosyntactic patterns in the Kiranti language Chintang. This language has many properties which make it especially interesting for this type of analysis (in context of language acquisition research).

Part I

Item-based patterns and language acquisition



A fractal pattern

1.1 Overview

The problem of language acquisition is inseparable from the following two questions:

1. What kind of a system is a language?
2. What is required to successfully learn such a system given a realistic set of conditions?

The first question deals with the complexity of the task to be learned. Clearly, this question is not trivial. Considerable difficulties start immediately after one tries to narrow this question down, due to the the semantic ambiguity of notions *language* and *system*. In following, under "language" as used above I will assume the speaker's apparent knowledge of the language (as a faculty of communication) structures. This knowledge allows him to recognize, comprehend and produce correctly formed utterances. This type of knowledge was coined by Chomsky (e.g. Chomsky, 1995) as *linguistic competence* (cf. also *langue* from Saussure et al., 1960)). It should not be confused to speakers *linguistic performance*, or the way an individual speaker uses his knowledge. That is, the speaker may fail to communicate successfully even if he possesses knowledge about the language, due to some performance impairments.

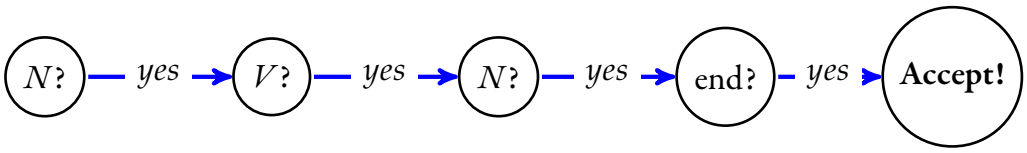
The notion of "system" is even more difficult to pinpoint. In order to avoid a discussion which would ultimately lead to the review of all literature on linguistic ever written, I will instead consider an associated notion of language's complexity (in a mathematical sense). Here, a language is simply a set of all grammatical utterances (and their meanings) and a language system is a black box containing a set of rules which describe the language. Then, the complexity is the mathematical notion which estimates the effort which is necessary to *decide* the language: that is, to test if any given utterance belongs to the language. The complexity of a system which consists of a small number of independent rules is lower than the complexity of a system with a high number of rules interacting with each other. To illustrate this notion, consider two simple languages consisting of one rule only which both are real subsets of English:

(1) **Utterances:**
Peter sees Paul
Natasha likes Peter
Rule:
$S \leftarrow N\ V\text{-s}\ N$

(2) **Examples:**
Paul, whom Peter sees
Paul, whom Peter, whom Natasha likes, sees
Paul, whom Peter, whom Natasha, who Alex watches, likes, sees
Rule:
$S \leftarrow N\ [\text{who}|\text{whom}\ N]^n\ V^n\ V$

The first language has only utterances which consist of three words which always come in a specific order. Imagine a hypothetical machine (Figure (2)) which consists of multiple states. Each state "eats" a word from an utterance and tests whether it matches a specified condition. If yes, the machine moves o the next state (and thus the next word), if no, the machine fails. Such machine is very simple, as it does not need to know which words came prior to the one currently being under inspection.

Figure 1.1: Deciding a simple language



However, when deciding the second language, one needs to keep track of the number and type of words which occurred prior to the current position. This is because the number of verbs has to match the number of nouns and this number is potentially unlimited. A simple machine, like (Figure (2)) cannot perform such task. Thus, the decision algorithm for this language requires more resources and therefore, the second language is more complex than the first one.

Of course, natural languages are even more complicated. It is generally assumed that most natural languages belong to the class of *context-free* languages, al-

though recent results suggest that this is not true for at least some languages (for more details, see Kracht, 2003; Shieber, 1985). Various formal theories of language grammar pursue to further capture the mathematical structure behind natural languages, thus narrowing down the question of complexity. While no definite answer has been found yet, the research results from this field were enough to trigger a vivid discussion in the literature.

This discussion concerns the question (2): what is required to successfully learn a language? Clearly, as only humans but no other known life forms are capable of learning a language, the learning devices must be something only humans possess. Some scholars claim that humans have genetical endowment, a kind of inborn knowledge concerning the language (linguistic nativism). Another argue that general cognitive abilities, which humans use for all kind of cognitive tasks, are sufficient to learn a language, with no innate language-specific knowledge required (linguistic empiricism).

In this chapter of my thesis, I want to give a brief overview on some chosen topics of this discussion. This overview in no means pretends to be exhaustive. My goal is to discuss some popular claims and results which ultimately outline the significance of item-based linguistic patterns for the research on language acquisition.

1.2 The innateness hypothesis and its criticism

1.2.1 Nativism vs. empiricism

A vivid discussion in the literature considers the question whether language is an innate human feature. Here, I follow (Pullum & Scholz, 2002) in adopting a dichotomy of possible views on this topic, which comprises the opposition between the *linguistic nativism* and its logical counterpart, which I will here refer to as *linguistic empiricism*[1]. In short, nativist position is that humans are born with innate linguistic knowledge (knowledge specific to the domain of language), while empiricists deny such claims. Technically, linguistic empiricism, as used here, may refer to a number of distinct views (and is therefore a potentially misleading term). For instance, a possible view is the one which denies innate human knowledge altogether, as opposed to the view which assumes some innate knowledge, but none language-specific. Both this views fall under the notion of empiricism in this discussion. The crucial difference between nativism and empiricism are the consequences for the language acquisition process: nativism presupposes the *innately-primed* learning mechanism, where children use their inborn linguistic knowledge to learn the language (and therefore, require no substantial language input to learn particular aspects of the language); while empiricism propose *data-driven* learning, where children learn the language solely from their experience, by inferring the language knowledge from the language input.

1.2.2 Poverty of the stimulus argument

Probably one of the most prominent supporters of linguistic nativism are Noam Chomsky and his followers. Chomsky's claims that children possess innate language knowledge are well known and currently belongs to the mainstream linguistic tradition. The basis of this claim is the *Poverty of the Stimulus* argument (cf. Chomsky, 1980b). In short, the argument is "you can't get from here to there": it is claimed that the information provided in the language input children receive is not enough to infer the correct generalizations about the syntax of the language. Pullum & Scholz (2002) summarize this argument in a following way:

[1] such dichotomy is clearly an oversimplification, but it is sufficient for the current discussion

(3) People attain knowledge of the structure of their language for which no evidence that is adequate to the task of learning this structure is available in the data to which they are exposed as children.[2]

Providing that the above statement is correct, there is an apparently paradox conflict between experience and attained knowledge (also known as *Plato's problem*). If the knowledge does not come from experience, it has to be innate, as the only viable alternative. Innate knowledge provides a solution for the Poverty of the Stimulus problem: indeed, if at least a part of language is innate, then it does not need to be learned in the first place.

The Poverty of the Stimulus Argument is central to the Chomskian tradition of generative syntax based on the Universal Grammar (UG). Chomsky first postulated the existence of the Language Acquisition Device — a hypothetical "organ" in a human brain which is responsible for learning the language. In a later version of his generative syntax theory, the Device has been replaced by the Principles and Parameters framework. The framework assumes the existence of a limited, innate set of possible syntactical rules (principles) which are a part of the human genome. The syntax of each human language ultimately builds upon these few principles. To account for the typological variance, the principles can be flexibly "tuned" by various switches (parameters). Enumeration of each possible parameters combination then describes the syntax of each possible human language. Therefore, language acquiring children do not need to learn the syntax from scratch — as it is sufficient to learn the corresponding parameters settings. One important consequence of this theory is the *continuity assumption* (Pinker, 1984): that the underlying linguistic competence does not differ between child and an adult. What differs is the language-specific knowledge: while an adult has mastered the lexicon and the individuals parameters of the language, the child is yet to do so.

However, the validity of the Poverty of the Stimulus argument is a topic of discussion. Many have argued that the very basic assumptions which constitute the argument are built on false premises. Below, I give a short review of these claims in an attempt to clarify the status of the Poverty of the Stimulus argument. In particular, as Pullum & Scholz (2002); Scholz & Pullum (2002) provide a very detailed overview and criticism on this topic, I will adopt their argumentation in parts.

[2] this is a weaker version of the statement found in in (Hornstein & Lightfoot, 1981)

Between the arguments which were made in favor of Poverty of Stimulus, two are particularly substantial. I refer to them as *a)* the *lack of positive evidence* argument and *b)* the *lack of the negative evidence* arguments.

1.2.2.1 Lack of positive evidence

The first argument (lack of positive evidence) reflects the claim the language input children receive does not contain enough sentences with complex grammatical structures and hence, provides no means for children to learn them. Popular example is auxiliary fronting in English polar interrogatives:

(4) a. (i) The man is smiling. (*declarative*)
(ii) Is the man _ smiling? (*interrogative*)
b. (i) The man who is standing there did smile. (*declarative*)
(ii) Did the man who is standing there _ smile? (*interrogative*)
(iii) *Is the man who _ standing there did smile? (*interrogative*)

The correct rule in formulating such interrogatives is to move the auxiliary verb of the main clause to the front. The last sentence is ungrammatical as not the main clause auxiliary, but the embedded clause auxiliary is fronted. This shows that not the linear order of the auxiliaries but the internal structure of the clause is important here.

The argument is that interrogatives of type (4-a) are frequent in the speech, while the ones of type (4-b)[3]: are very seldom or even do not occur at all, as in Chomsky (1980a, pp. 40)

> A person might go through much or all of his life without ever having being exposed to relevant evidence, but he will nevertheless unerringly employ [the correct rule] on the first relevant occasion

If this is true, and only the interrogatives of type (4-a), then — following Chomsky — children should infer a simpler (and wrong) generalization *move the first auxiliary to the front*, but not the correct one *move the main clause auxiliary to the front*. The children, however, do not make such mistakes (see S. & M., 1987; Crain, 1991). Therefore, at this point children must already have some knowledge which causes them to use structure-dependent rules instead of linear-order-dependent

[3]Pullum uses the name "Chomsky-sentences" in his internet log entry (http://itre.cis.upenn.edu/~myl/languagelog/archives/000156.html)

rules. Chomsky's conclusion is that learning children impose natural restrictions on the domain of syntactic operations — which has to be innate.

This argument has two weaknesses. First, the assumption that sentences like (4-b) do not occur in the speech is not backed up by data, but was proposed by Chomsky as a product of his intuition. As Pullum & Scholz (2002) point out, a simple corpus search reveals that the frequency of such sentences in the speech is higher than Chomsky has suspected. Second, it is possible that children independently learn that structure-dependent representation is preferable — by learning other constructions or following some other cognitive principles[4]. This preferences lead them to the correct generalization of the auxiliary fronting, even if there was not enough input — i.e. under such conditions the structure-dependent rule would actually be the simpler one. A. Perfors & Regier (2006) trained a simple neuronal network on a English corpus which contained no complex sentences. Given the choice between a flat and a structure-dependent grammar representation, the network preferred the last one. In another words, the network could infer that the language is structure-dependent without actually having any direct evidence for it. It is possible that a child could use similar strategy, which does not require innate language-specific knowledge.

1.2.2.2 Lack of negative evidence

The second argument (lack of negative evidence) revolts around the fact, that children are seldom given feedback about the grammaticality of the sentences they produce. Some have claimed that such type of evidence (negative evidence) is essential to avoid overgeneralization of the inferred rules (see Sokolov & Snow, 1994, for overview). This claim is further backed up by the classical result from the learnability theory (Gold, 1967). Gold studied a number of formal language learners using the notion of *learnability in limit* — a language was considered learnable if the learner was able to solve the language decision problem after being exposed to a finite number of sentences. He showed that positive evidence alone was not sufficient to learn even some simpler classes of formal languages. On the other hand, any language was learnable if the learner had access to negative evidence.

A nativist solution to this problem is that the innate language knowledge constrains the possibilities of rule representation, such that no rule overgeneration is

[4] In particularly, humans apparently show general preferences for hierarchical representation of information (part-of relations) outside the language, in the visual, auditory and other domains.

possible. Thus, the inborn grammar provides the required negative evidence indirectly in no contradiction to Gold's results. Data-driven learning can only rely on positive evidence and thus is not adequate.

There are some objections against this argument. For instance, in Bowerman (1988); Bohannon & Stanowicz (1988) it is shown that adults react to children's errors in a specific and systematic way. It is argued that such response could indirectly provide children with the negative evidence necessary to successfully acquire a language. This suggestion was criticized by Gordon (1990), who claims that this kind of negative evidence was neither accurate enough nor sufficient to allow learning in Gold's (1967) sense.

However, new results in learnability theory suggest that Gold's results may be overrated. So, Shinohara (1994) showed that a rich subset of context-sensitive languages can in fact be learned from the positive evidence alone (see Scholz & Pullum, 2002, for discussion). Also, Gold's proofs are based on the assumption that the sentences are presented in a random order, which is not exactly true for natural languages. In a flow of speech, the order, content and structure of the sentences are not only determined by the grammar of the language, but also subject to numerous constraints placed by by semantics, pragmatics, context etc. Basically, there is a "hidden" structure to the language which a traditional grammar does not capture. In short, language is a *stochastic process*. There is evidence that such "hidden" stochastic structure can beneficially affect the language learning. So, Rohde & Plaut (1999) show that a connectionist network achieves better results when trained on a corpus of semantically constrained pseudo-English; their simulation the network was able to acquire relatively complex syntax of nested relative clauses. The network had considerable difficulties with language learning when no semantic constraints were present.

1.2.3 Intermediate summary

The above arguments suggest that evidence in favor of the Poverty of the Stimulus (and language-specific innate knowledge) is weaker than assumed by many scholars. It is important to note though, that they do not mark linguistic nativism as wrong; rather, they restore the "balance of power" between nativism and empiricism (which was traditionally disdained). The intermediate conclusion is that the theories of linguistic nativism found in the literature seems to be more often motivated by the intuition then by the factual evidence.

One aspect of language which has been overlooked by the proponents of linguistic nativism is the stochastic nature of the language (as I have pointed out above). This aspect is nevertheless crucial for the success of an empiricist account. Stochastic regularities in the language input provide an additional source of information which can be used by the children.

A logical way to continue the research is to further pursue the question of what — and under which conditions — can be learned; therefore, to further explore the possibilities of the data-driven learning while considering the particularities of the language input of the children (Pullum & Scholz, 2002, arrive to the similar conclusion). One highly "affordable" method involves corpus linguistic. Real-life corpora of actual children's language input provides an adequate way to study the stochastic properties of the languages — and may shed some light on details of how a language may be acquired. This has long been recognized by many scholars, considering the recently increased academic interest in this particular direction.

1.3 Data-driven learning and the acquisition of lexical categories

1.3.1 Cues in the acquisition of lexical categories

Any theory of language acquisition — be it based in linguistic nativism or linguistic empiricism — must provide some account of data-based learning. Even when one assumes that children have innate knowledge about syntactic representations and lexical categories, they still have to map this knowledge to the actual constructions and lexemes in the language. Of course, no one (I dearly hope so) will assume that the lexicon is innate — that is, that individual lexemes of the language are part of the human genes. Even if the child has an inborn knowledge of the category **noun**, this knowledge does not provide it with the list of the actual nouns in the language. Rather, the child has to learn the nouns from the language input. In this regard, both nativist and empiricist theories start the learning process with a similar set of conditions.

The learning process is further complicated by the fact that languages may differ substantially in the numbers and types of lexical categories they exhibit and in the ways the constructions are coded. A single lexical category may manifest multiple formal subclasses (inflectional classes) or exhibit considerable structural complexity (verbs in polysynthetic languages). Thus, an adequate learning mechanism has to be flexible enough to account for such variation and still robust enough to result in a correct classification.

A number of accounts have been proposed which describe possible approaches to lexical category acquisition. They are all based on the fundamental observation that grammatical structures and their realization are isomorphic — that is, similar structures are realized in a similar way (and via versa). The notion of *realization* here includes a wide range of properties. It is clearly outside the scope of my thesis to discuss all such proposals in detail, but I will give a very sketchy and incomplete overview of some prominent proposals.

One cue to categorization is the phonology. It has been observed that different lexical categories may have different phonological properties. For instance, English nouns tend to have more syllables than English verbs, they also show some systematic differences in regards to stress (Kelly, 1992, 1996). Shi et al. (1998) discusses

some phonological properties relevant to lexical category distinction in Turkish and Mandarin Chinese.

Semantics is another cue which may be relevant in acquisition of lexical categories. Different lexical categories often have a semantic prototype. Virtually every language differentiates between *entity* and *action* in grammar, which gives rise to categories of nouns and verbs. (Pinker, 1984) proposed a semantic bootstrapping hypothesis — a nativist approach, where semantic cues are initially used to identify innately-motivated lexical categories (Grimshaw, 1981, c.f). Another account is provided by Tomasello (2003). There, it is proposed that the children identify communicative functions associated with individual words and categorize the words accordingly.

Yet another source of categorization information, which has been shown to be especially powerful is the *distributional information*.

1.3.2 Distributional learning

Distribution of an item is understood as a sum of all environments where it occurs. When studying the distributions of different items, particular patterns emerge, which can be used to discriminate the items from each other. This particular property of language has been long exploited by linguists to study phonemes and morphemes in various languages. This approach was strongly influenced by structuralists (Harris, 1964, cf.).

Associated with this is the idea of *distributional learning*, a learning strategy which makes use of the distributional information to learn particular structures. A number of studies show that humans have powerful pattern discrimination tools on their disposal. Even less than a year young infants are able to discriminate regularities in the language input, both for natural (Jusczyk, 1997; Gerken et al., 2005) and artificial (Saffran et al., 1996; Gomez & L., 1999) grammars. The later are especially notable, as they show that children are capable of recognizing speech patterns even after a minimal exposure. In this experiments, children correctly performed grammatical judgement and word segmentation tasks after only two minutes of exposure to an artificial grammar which they have never learned before.

It has also been suggested that distributional learning can be used to learn lexical categories. For instance, Maratsos & Chalkley (1980) note that words of the same category have similar morphological environment, e.g. stems which take suffixes *-ed* and *-s* are usually verbs. Approaches based on word co-occurrence statistics

(Redington et al., 1998; Mintz, 2003) showed that distributional analysis alone results in highly accurate clustering of the lexical categories. An early study used a spearing-activation network model (Kiss, 1973). Yet another approach based on heuristic learning Cartwright & Brent (1997) obtains similar results. Mintz (2002) shows that adults can use distributional information to discriminate categories in an artificial language. All these results suggest that distributional information alone (with no phonological / semantical cues) makes good predictions of lexical categories. When combined with phonological cues, the prediction accuracy is further improved (Monaghan et al., 2007).

Pinker (1984, 1987) objects against the distributional learning as a main mechanism for category acquisition. His argument is that a large-scale distributional analysis would require considerable resources as the number of combinatorial possibilities is very large. Another one is that "pure" distributional learning could result in wrong categorization — as in an example provided by Pinker:

(5) John can fish
John ate fish
John ate rabbit
⇒ *John can rabbit

An objection to Pinker's arguments is that distributional learning does not involve brute-force enumeration of all statistical information but rather uses a more intelligent approach. Crucial distinction here is between the notions of *statistic-based* learning as opposed to *statistic-driven* learning (Elman, 2002). Statistic-based learning involves, as the name may suggest, learning of statistics, while statistic-driven learning uses aspects of stochastic information to infer knowledge about the language. For instance, Mintz (2003) shows that lexical categories can be predicted with high accuracy only considering a relatively small number of lexical frames like *I [...] it*, *put [...] in*. This leads to the assumption that very frequent distributional patterns may provide the large share of required information and the learning mechanism of the child has means to filter such patterns. This way, the combinatorial explosion Pinker talks about can be prevented. The example (5) is particularly weak, as cases of noun / verb homophones are hardly frequent enough to irritate a distributional learning mechanism which relies of statistics.

1.3.3 Possibilities of data-driven learning

Above I outlined the discussion on how lexical categories may be acquired by a learning child. The results allow to make two conclusions: a) acquisition of lexical categories involves data-driven learning and b) acquisition of lexical categories may involve different sources of information: semantical, phonological and distributional. Overall, it is likely that the child's learning mechanism uses these sources simultaneously and according to their relevance. That is, the learning is *optimal* in the sense that it uses the information which is most likely to lead to success. Thus, children can use powerful and flexible pattern discrimination methods which strongly rely on the "hidden" stochastic structure of the language.

Recently, a number of data-driven approaches showed that even with no innate knowledge, fairly complex structures can be learned. As already noted, A. Perfors & Regier (2006) argue that the learner should be able to recognize the structure-driven nature of the language using data-driven algorithm only. An interesting suggestion was made by Elman (1993), who shows that a connectionist network is able to acquire a fairly complex grammar if it's capabilities are limited at first and advance gradually. In the initial phase, where the network has very limited resources, it is only able to acquire very simple structures. However, as the capabilities of the network grow, the knowledge about the simple structures allow it to learn more complex structures. One may assume that children behave similarly, acquiring more complex structures as their cognitive ability develops. Other studies with connectionist networks (Rohde & Plaut, 1999; Elman, 2002) also suggest that data-driven learning is a powerful method capable of more than claimed by many language nativists.

1.4 Frequent patterns in the language

1.4.1 Frequent patterns in the child language

It is a known result of language acquisition research that the early child language syntax revolts around rigid item-based[5] templates, or *frames*. So, Tomasello (2003) compiled a database following the development of his daughter's speech. He observed that many verbs were used in a very small number of unique frames, where the frame use was verb-specific. Similar was also observed for morphological marking. Lieven et al. (1997) studied the language of English speaking children. They found out that virtually all (92% on average) utterances, which contained more than one word, produced by the children followed only 25 templates. One additional observation was that the templates were child-specific, i.e. they differed from child to child.

These results suggest that at this stage children do not have much abstract knowledge about the syntax of the language. Rather, they first learn usage-specific templates like *Where is X?*, *This is X*, *Take the X*. These observations cast shadow's of doubt on Pinker's (1984) continuity assumption and on the language nativism (at least in its mainstream form) in general.

Of course, it is clear that at some point of time the children abandon these templates and move to more general syntactical rules. An empiricist theory of language acquisition must provide an explanation for such process. One initial proposal is made by Abbot-Smith & Tomasello (2006). They suggest that the usage-specific templates evolve by the faculty of abstraction under the influence of semantical and pragmatical knowledge. This is a quantity-to-quality approach, as opposed to the quality-to-quantity approach of linguistic nativists. For instance, question templates *Who is X?*, *What does X do?*, *Where is X?*, *What did X do?* are summarized as *WH AUX X ...*, which is the basic syntax of an English interrogative. Of course, this suggestion does not provide much more beyond a proposal, let alone a complete and working theory of data-driven syntax acquisition. However, it will hopefully trigger a branch of interesting theoretical research which further explore the possibilities of such proposal.

[5] *item* here refers to words and/or morphemes

1.4.2 Frequent patterns in the child-directed language

So where do these frames come from? Apparently, an empiricist account will state that children learn them by mimicking the information they get, that is, the frames have to originate from the language input. In following, I will discuss some recent studies which have dealt with the structure of child-directed speech.

Using a trigram co-occurrence analysis, Mintz (2003) identified a small number of frequent templates of form *X [...] Y*. They observed that virtually all words which could occupy the middle position of such templates were of the same lexical category. Moreover, the template edges (*X, Y*) often belonged to a closed lexical category.

Cameron-Faulkner et al. (2003) analyzed the beginnings of utterances in English child-directed speech, based on recorded mother-child interactions of twelve English-speaking mothers. They compiled the frequency of the utterance-initial sequences which contained up to three words (morphemes). The study showed that a restricted set of patterns accounted for a large amount of all child-directed utterances. So, 51% of all utterances produced began with one of 52 patterns, 45% began with one of 17 words.

However, one may argue that the high degree of lexical restrictiveness in the sentence-initial position may be an artifact of English grammar. English is known to have a very rigid word order, with virtually no inflectional morphology. In addition, the language has determiners, an obligatory copula and relies heavily on use of auxiliaries. It is possible that a language with a more "free" grammar and rich morphology, which allows grater variation within the syntax, will also show significantly less repetition in utterance-initial position.

To test this, Stoll et al. (2009) conducted a study on child-directed speech in three typologically different languages: Russian, German and English. Russian was chosen as a language which (at least in part) exhibits opposite traits to English: the word order in Russian is relatively free and depends on pragmatics, it has rich inflectional morphology, no determiners and no obligatory copula or auxiliaries in present tense. Most specifically, Stoll et al. counted *core frames* — frequent frames which were present in the speech of at least 50% of mothers.

A limited number of frequent utterance-initial frames were found in all three languages. There were statistically significant differences between the languages in respect to the number and the length of frames. So, English had the highest number of frames, which also constituted of the largest number of morphemes.

On the other side, Russian had the lowest number of frames which were also the shortest — usually only one morpheme in length. German scored a place between English and Russian. All this can be explained by the typological properties of the considered languages[6] (see Stoll et al., 2009, for detailed discussion). These results suggest that there is at least some truth to the above hypothesis that lexical restrictiveness depends on the properties of the language grammar.

However, there were also considerable similarities between the three languages in regards to the utterance-initial lexical restrictiveness. As in the Cameron-Faulkner et al.'s (2003) study, a large proportion of all utterances began with one of the frequent frames — 64% of all utterances in English, 63% in German and 53% in Russian (with 122, 79 and 63 frames respectively). Also, the breakdown of the most frequent utterance-initial words was similar in all three languages and included large number of pronouns, wh-particles, imperative verbs and demonstratives.

In the final part of my thesis I do a similar study on the Kiranti language Chintang. In regards to morphological complexity and word order freedom Chintang is more "extreme" than Russian. In addition, it features massive argument drop. The frequent pattern analysis revealed a picture which overall closely resembled the findings of Stoll et al., which is particularly astonishing, as Chintang is typologically and culturally very distant from the European languages.

[6]The high number of frames in English was also in part due to the counting method. Stoll et al. counted frames which contained another frames separately. Thus, *That*, *That is* and *That is a* were three different frames.

1.5 Outlook

The above results suggest that frequent item-based frames play an important role in language acquisition. They constitute a prominent portion of the stochastic information which is contained in the speech and thus may be easily detected and used by statistically-driven learning mechanisms.

For the language acquisition research this means to further focus attention on the study of frequent frames, both in the language produced by children and in the child-directed language. The goal of such studies will be to *a)* determine which constructions can be learned from the frequent frames in the child-directed language, *b)* which frames are used by the children in their speech and *c)* what is the correlation between the frames in the child's speech and the child-directed speech. Also, this research should be preferably carried out with typologically different languages to study what are the differences similarities in the acquisition of the grammars of different types are.

This leaves open the question about the methodology of frequent frame recognition. Previous studies used various approaches to find frequent frames, but none of them was flexible. That is, they could only find frames of a particular type; for instance, continuos frames in the beginning of the utterance as in (Stoll et al., 2009; Cameron-Faulkner et al., 2003). However, it is not obvious that the set of relevant frames is that constrained. A generalized method which imposes few or no restrictions on the pattern shape — that is, which is able to extract *all* frequent morphosyntactic patterns from the speech data — would be a valuable tool at a language acquisition researcher's disposal.

Part II

Identification of frequent patterns in corpora

By courtesy of Randall Munroe, http://xkcd.com/

2.1 Overview

Despite the apparent interest in the study of frequent patterns in speech, no general method to identify such patterns has been proposed so far. Previous studies devised hand-tailored methods suitable for their purpose. So, in order to study the degree of utterance-initial lexical restrictiveness Stoll et al. (2009); Cameron-Faulkner et al. (2003) compiled the information about frequent uninterrupted word sequences, beginning with the first word of the utterance. In category acquisition studies, Mintz (2003) collected frequent word pairs which encircled another word, that is, frequent templates of the form *X* ... *Y*. Redington et al. (1998) computed bigram word co-occurrence statistics. Cartwright & Brent (1997) used a heuristic-driven category learning algorithm, effectively bypassing the need of frequent frame analysis.

The obvious disadvantage of the hand-tailored methods of pattern identification is that they can only find the patterns they are designed to find. They place a burden of predicting how the interesting patters may look like onto the researcher. If a pattern is relevant to the study, but slightly deviates from the predicted form, it won't be detected. Another disadvantage of hand-tailored methods is that they — as suggested by their definition — have to be individually crafted for each study, which is additional work. Furthermore, extension of such methods to account for additional pattern types is often cumbersome.

In this section I describe a general framework of frequent pattern identification. The framework searches for patterns within a set of linear sequences of elements, such as words in the utterances of a language corpus. It can detect a large variety of patterns, independent of their position in the utterance or the availability of (multiple) gaps.

The core of the framework is the theory of formal languages and regular expressions. Regular expressions have long been known in computer science as a powerful and convenient tool for pattern matching. In a nutshell, a regular expression describes a particular set of linear item sequences. Such sequences could be a text string (sequence of letters), words in a sentence etc. The regular expressions are written in a special language with well-defined syntax and semantics. They are widely used by many computer specialists which work with textual data, as regular

expressions provide fast and flexible searches in large texts. For this reason, they are becoming increasingly popular with corpus linguists.

The usual case involves a situation when a regular expression is known — that is, the user wants to test if a sequence follows a specific pattern. For instance, a text editor could search for all instances of particular phrase, a web site programmer would want to test whether the entered email addresses are formated correctly, and a corpus linguist could look for all nouns with the suffix -ɲa. Therefore, the common application of regular expressions ultimately boils down to answering the question *does a particular regular expression describe a particular sequence?* A large number of robust algorithms have been developed for this purpose. Today, usage of regular expressions for sequence matching is lightning fast and requires very few resources. I will not discuss such algorithms here as they are clearly outside the scope of my thesis.

However, when one is concerned in finding patterns in text, the reverse question is of great interest: *which regular expressions describe a particular sequence?*. Clearly, if a practical answer to this question exists, a whole new approach to pattern identification in corpora emerges. As regular expressions are a natural choice to describe patterns in languages, the idea would be to find *all* regular expressions which can describe at least one utterance in the corpus and then filter these regular expression in respect to the number of total utterances they derive. In the end, one would obtain the list of frequent patterns.

In this chapter of my thesis, I introduce some basic notions from the theory of formal languages and explore the possibilities of solving the above question. To my knowledge, no prior solution to this problem has been attempted.

2.2 Patterns in formal languages

2.2.1 Preliminaries

A set of symbols Σ is called an *alphabet*. A sequence of symbols from Σ is called a *string* (or a word) over Σ. A special string is the *empty* string $\varnothing$. Two words can be *concatenated* (placed after each other in a sequence) to produce a new string. Concatenation with the empty string does not change the original string. In following, I will write concatenation simply as juxtaposition.

The *Kleene's closure* X^* over the set of words X is defined as the smallest set which contains X and is closed in respect to the string concatenation, that is, a concatenation of each two strings from X^* also is in X^*.

Thus, $(\Sigma \cup \varnothing)^*$ is the set of all possible strings over the alphabet Σ, including the empty string. Each subset of $(\Sigma \cup \varnothing)^*$ is called a *formal language*. A language is finite, if the number of the strings it contains is finite, otherwise it is infinite.

A formal language is described by a *formal grammar*. There are different approaches to formal grammar representation, with one of the most popular ones being the generative grammars, proposed by Chomsky (1956) . A classic formalization describes a formal grammar as a tuple $G = (N, \Sigma, P, S)$ which includes:

- A finite set N of *nonterminals*
- A finite set Σ of *terminals* ($\Sigma \cap N = \varnothing$) (also *alphabet*)
- A finite set of *production rules* P, where each rule is a mapping in form of

$$(\Sigma \cup N)^* N (\Sigma \cup N)^* \rightarrow (\Sigma \cup N)^*$$

- A *start symbol* $S \in N$

The grammar generates all strings in a formal language in a following way. First, the final language only includes strings which contain symbols from Σ (no nonterminal symbols). As can be seen from the definition, a production rule transforms a string composed of at least one nonterminal and zero or more terminal symbols, to a string which is composed of terminal and nonterminal symbols (which may be empty as well). The grammar applies the production rules to generate new strings

by rewriting suitable parts of already generated strings, starting with the string S. After all possible string rewriting has been performed (which is potentially an infinite process), the strings which contain only terminal symbols are exactly the strings of the language described by a grammar.

As an example, let us consider a language which consists of strings *ab, aabb, aaabbb*, that is, $L = a^n b^n$, where exponentiation means n-time concatenation. This language can be generated by a following grammar:

- $\mathrm{N} ::= \{\mathrm{S}, X\}$
- $\Sigma ::= \{a, b\}$
- $P ::=$

$$\left\{ \begin{array}{lcl} S & \Rightarrow & aXb \\ X & \Rightarrow & aXb \\ X & \Rightarrow & \varnothing \end{array} \right\}$$

It is easy to see that this grammar indeed generates L. Starting with the string S the grammar expands it to aXa. Here, X can be either rewritten as aXb, or as an empty string. Overall, the results are sequences of some a's followed by the same number of b's, which constitute exactly the strings from L.

Natural languages can be described using formal languages. Formal languages describe arbitrary sequences of elements, while natural languages clearly are such sequences. As a matter of fact, much of the initial theory of formal languages was established in order to find adequate models of natural language representation. While the research in this area still has to struggle with considerable difficulties, the formal language theory is very well applicable to language corpora, as I will try to convince the reader by the end of this chapter.

2.2.2 Regular languages

A class of formal languages with particular properties are the *regular languages*, described as follows:

1. The language $\{\varnothing\}$ is a regular language
2. For an alphabet Σ, languages $\{\{a\} \,|a \in \Sigma\}$ are regular

3. If L_1 and L_2 are regular languages, languages $L = L_1 \cup L_2$ and $L = \{ab | a \in L_1, b \in L_2\}$ are regular

4. If L is a regular language, then L^* is regular

5. No other language is regular

A generative grammar which describes regular languages is one which only has production rules of two kinds:

$$\begin{aligned} A &\Rightarrow a \\ A &\Rightarrow aB \end{aligned}$$

where A, B are nonterminal symbols and a is a terminal symbol (also including the empty string). It is not difficult to see that production rules generate exactly the languages for which the above description applies. The $A \Rightarrow a$ type of rule describes all one-symbol regular languages and $A \Rightarrow aB$ type of rule accounts for regular language concatenation. The grammar of the union of two regular languages is simply the union of their grammar components (symbols and production rules)[7]. The grammar of L^* over a regular language L is constructed by replacing all rules $A \Rightarrow a$ with $A \Rightarrow a\mathrm{S}$ where S is the starting symbol in the grammar generating L.

The regular languages are prominent, because they form one of the simplest formal languages known. In particular, regular languages have no unlimited memory for inter-string dependencies, that is, they cannot derive languages like $a^n b^n$. This also means that natural languages are not regular, as this pattern can be found there. Example from English:

(6) The dog, [who the man, [who the girl, [who the cat saw], talked to], bought].
$\mathrm{NP}^n\ \mathrm{V}^n$

However, due to their simplicity, regular languages are easy to describe and to deal with. Even if they cannot describe the whole range $a^n b^n$, they still can describe finite strings like *ab*, *aabb*, *aaabbb*, etc. — which in many cases is an adequate approximation. This is further confirmed by the following lemma.

[7] Of course, the set of nonterminals in both languages should not overlap. This can be easily achieved by item renaming

Lemma All finite languages are regular

Proof. Each finite language L over the alphabet Σ can be represented as a union set of languages S_i, where each S_i contains one string of L. Each string is a finite sequence of symbols from Σ can be in its turn represented as a concatenation of regular languages A_j, where each A_j contains one-symbol strings from Σ. Therefore, S_i are regular and so is L. □

Text corpora, due to their finite nature, can be thus sufficiently described with regular languages. This significantly reduces the complexity of the task at hand, while still retaining the ability to recognize all patterns.

2.2.3 Regular expressions

A regular expression is a string of a special meta-language which describes regular languages. More specifically, a regular expression r over an alphabet Σ is a string that belongs to the language $Regex(\Sigma)$ over the alphabet $\Sigma \cup \{+, *\}$. To distinguish regular expressions from other strings I will enclose them in edge brackets $\langle\rangle$. The syntax of this language is defined as follows:

1. $\langle\varnothing\rangle$ is a regular expression (equivalent to empty regular expression $\langle\rangle$)
2. Strings $\langle a\rangle | a \in \Sigma$ are regular expressions
3. If $\langle r\rangle, \langle q\rangle$ are regular expressions then $\langle r + q\rangle$ (union), $\langle rq\rangle$ (concatenation) and $\langle r*\rangle$ (Kleene's closure) are regular expressions
4. Nothing else is a regular expression

A regular expression is used to describe a language over Σ. This is made possible by adopting clear semantics for $Regex(\Sigma)$:

1. The regular expressions $\langle\varnothing\rangle$ and $\langle a\rangle$ ($a \in \Sigma$) describe the empty string and the string a, respectively
2. The regular expression $\langle r + q\rangle$ (where $\langle r\rangle, \langle q\rangle$ are regular expressions) describe either the string which is described by $\langle r\rangle$ or the string which is described by $\langle q\rangle$

3. The regular expression $\langle rq \rangle$ (where $\langle r \rangle$, $\langle q \rangle$ are regular expressions) describes a string which contains a string described by $\langle r \rangle$ followed by a the string described by $\langle q \rangle$ (concatenation)

4. The regular expression $\langle r^* \rangle$ describes all strings which belong to the Kleene closure of the strings described by $\langle r \rangle$

If a regular expression $\langle r \rangle$ describes a string s we say that $\langle r \rangle$ *matches* s. A *trivial regular expression* which matches a string s is simply the regular expression string $\langle s \rangle$.

There is a clear one-to-one mapping from the regular expression syntax to the definition of the regular languages in the previous section. Here, $+$ corresponds to language union, regular pattern concatenation corresponds to regular language concatenation and $*$ corresponds to the Kleene closure of the regular language. Thus, regular expressions sufficiently describe all regular languages.

A number of writing conventions, similar to that used in arithmetics, have been adopted for representing regular expressions. The parentheses are used to group parts of the expression together. For instance, $\langle a(b^*)(a + (b + c)^*)e \rangle$ matches strings which begin with a, followed either by a or an arbitrary long mixed sequences of b and c, and end with e. For instance, matched strings are *abe*, *aae* but not *aabe* or *aba*. To reduce the number of required parentheses, operation precedence rules are introduced: $*$ has precedence over the concatenation and union $+$ has the lowest precedence. Therefore, the above regular expression can also be written as $\langle ab^*(a + (b + c)^*)e \rangle$.

2.2.4 Regular expression generation

Given a string s over a alphabet Σ, the task is to generate all regular expressions that match s. Formally, the task is to find a mapping $regex_\Sigma(s) ::= \Sigma^* \rightarrow \{\langle r \rangle | \langle r \rangle \in Regex(\Sigma)$ matches s.

Theorem 2.2.1. ***Construction of regular expressions***

Given a string s, every regular expression that matches s can be recursively constructed from the trivial regular expression $\langle s \rangle$ using only the following rules

1. Empty Concatenation

If $\langle r \rangle$ matches s, so does $\langle rq \rangle$ and $\langle qr \rangle$ where $\langle q \rangle$ is a regular expression which matches the empty string.

2. Substring Closure

   If $\langle rq \rangle$ matches s, so does $\langle (r + q)^* \rangle$. Similar, if $\langle (r + q)^* \rangle$ matches s then one of $\langle r \rangle$, $\langle q \rangle$, $\langle rq \rangle$, $\langle qr \rangle$ matches s

3. Extension

   If $\langle r \rangle$ matches s, so does $\langle r + q \rangle$, where $\langle q \rangle$ is an arbitrary regular expression. The reverse obviously applies.

From this three basic construction rules, a number of derived rules follow. For example:

1. If $\langle r \rangle$ matches s, so does $\langle r^* \rangle$ (Substring Closure and Empty Concatenation)
2. If $\langle rq \rangle$ matches s, so does $\langle (r + p_1)(q + p_2) \rangle$, $\langle (r + p_1)^*(q + p_2)^* \rangle$ (Extension and Substring Closure)

Proof. To prove the theorem, I make use of the fact that there is only a small number of distinct logical possibilities to construct a complex regular expression $\langle r \rangle$. We say, $\langle r \rangle$ *is reducible to* $\langle q \rangle$ if $\langle r \rangle$ can be constructed from $\langle q \rangle$ by application of the above rules.

The proof itself is based on mathematical induction.

First, assume that the length of s is one (thus, the string contains only one symbol). Let $\langle r \rangle$ be a regular expression which matches s. Following possibilities exist:

1. $\langle r \rangle = \langle s \rangle$ (trivial)
2. $\langle r \rangle = \langle pq \rangle$ — then either $\langle p \rangle$ or $\langle q \rangle$ must match the empty string (as $\langle r \rangle$ matches one symbol only) and the other one matches s. (Empty Concatenation)
3. $\langle r \rangle = \langle p^* \rangle$ — $\langle p \rangle$ must match s, as the string contains only one symbol (Substring Closure and Empty Concatenation)
4. $\langle r \rangle = \langle p + q \rangle$ — $\langle p \rangle$ or/and $\langle q \rangle$ must match s (Extension)
5. No other logical possibility exist

Regardless of $\langle r \rangle$'s form, it can be reduced to a less complex regular expression which matches the one-symbol s by inverse-application of the above rules. If one applies this reduction recursively, $\langle r \rangle$ — due to its finiteness — can be ultimately reduced to a regular expression containing one symbol only. Trivially, this must be $\langle s \rangle$. Thus, the theorem holds for all s with length of one.

Let us now assume that the theorem holds for any s with length less than n and consider the case where s's length is n. Again, considering the logical possibilities for $\langle r \rangle$ which matches s:

1. $\langle r \rangle = \langle pq \rangle$. Disregarding the trivial case where $\langle p \rangle$ matches the empty string and $\langle q \rangle$ matches s (or the other way round), $\langle p \rangle$ will match the first part of the string s and $\langle q \rangle$ will match the remaining part. Both these substrings have length less than n. Thus, $\langle p \rangle$ and $\langle q \rangle$ are reducible to the trivial form and so is $\langle r \rangle$.

2. $\langle r \rangle = \langle p^* \rangle$. Either $\langle p \rangle$ matches s or $\langle p \rangle = \langle p_1 + p_2 + \cdots + p_m \rangle$ so that the regular expression $\langle p_{i_1} \cdots p_{i_n} \rangle$ (for suitable i) matches s (Substring Closure / Extension). When a single $\langle p_i \rangle$ matches s entirely, we continue the argument from the top. Otherwise, each $\langle p_i \rangle$ will match a substring of s and thus be reducible to a trivial form (due to the length less than n). Hence, $\langle r \rangle$ can be reduced to $\langle s \rangle$.

3. $\langle r \rangle = \langle p + q \rangle$. Either $\langle p \rangle$ or $\langle q \rangle$ match s. (Extension)

4. No other logical possibility exists

The same reasoning as above applies here too: we can subsequently reduce $\langle r \rangle$ to the trivial form, by considering respective substrings of $\langle r \rangle$ recursively.

Using mathematical induction, the theorem is proven. □

2.3 Restricting the pattern language

2.3.1 Interesting and uninteresting patterns

The above Theorem 2.2.1 provides all the means necessary to generate all regular expression which match a string s. A straightforward algorithm is

$A \leftarrow \{\langle s \rangle\}$
loop
  $\langle r \rangle \leftarrow$ a regular expression which can be constructed from any $\langle r \rangle \in A$ using the derivation rules
  $A \leftarrow A \cup \{\langle r \rangle\}$
end loop

Still, this is far away from a practical solution. The number of regular expressions which match a single string is not finite — the Extension Rule alone grants it. Obviously, the above algorithm is only of limited use, as it will never terminate. From this I conclude that a practical implementation must somehow constrain the regular expressions it generates.

Fortunately, only a particular type of regular expressions are of potential interest to the research of frequent patterns in language corpora. In particular, the regular expressions should match homogeneous strings. Especially when researching the child-directed speech, the patterns should be as rigid as possible: a pattern which matches material too heterogeneous has no abstraction power and hence cannot be used to learn structural properties of the language.

For example, in a language with nominal determiners, a pattern of particular interest would be $\langle$ *[Det] [N]* $\rangle$. This pattern describes an important distributional property of nouns: they are often preceded by a determiner. As the number of determiners (closed word class) is far less than the number of nouns, such pattern provides a fixed frame which can be used to predict that the next word is a following noun.

On contrary, a pattern like $\langle$ *([Det]+[N])** $\rangle$, constructed using the Substring Closure rule, is of no interest at all, as it clearly overgenerates, describing strings which do not occur in the language. This pattern does not describe any existing distributional properties for either determiners or nouns. Similar reasoning applies to

a pattern ⟨ *([Det] [N])+[V]* ⟩ (Extension rule). Here, the pattern is not interesting as it matches heterogeneous constructions (either noun phrases or verbs) and thus provides no interesting distributional information — aside from a rather obvious observation that the language may contain noun phrases and verbs. Another trivial case of an uninteresting pattern is ⟨ *[...]** ⟩ (where *[...]* matches any item) — which basically provides no information at all.

2.4 Constraining the pattern derivation rules

Clearly, one obvious source of heterogeneity is the union operation $+$, which introduces optionality. In particular, the Extension rule is only of limited interest, as it easily overgenerates. Specifically, extensions of type $\langle r + q \rangle$ (derived from $\langle r \rangle$) should be avoided altogether, as such patterns either reduce the significance of distributional information carried or predict strings which are not in the language. In avoiding the union operation we also solve the problem of the pattern infinite variability, thus, we only have to deal with finite number of interesting patterns.

However, the union operation does have one particular relevant application. It can be used to describe groups of tokens of similar category (I will use square brackets *[]* to distinguish categories from the atomic tokens). Following reasoning applies: if a token from a given category occurs in a pattern, it is possible that another token from the same category may occur at the the same location in this pattern. For instance, for English, one would want to predefine *[Det]* as $\langle$ *the* + *a* $\rangle$. Then, given the string *the dog*, an appropriate algorithm would generate patterns $\langle$ *the dog* $\rangle$ and $\langle$ *[Det] dog* $\rangle$.

A straightforward application of the grouping mechanism is the introduction of the special *[...]* symbol which can match any token. This is important, as — combined with repetition (see below) — it allows for precise identification of fixed frames by exclusion of flexible material. For instance, *the [...] dog* matches any phrase which has an additional token between *the* and *dog*.

The Substring Closure derivation rule can be safely ignored, as it includes the union operation and thus introduces unwanted heterogeneity. Still, Kleene closure is of interest, as it provides the faculty of repetition, allowing us to match arbitrary sequences of particular item. Thus, I redefine the * operation as *match a token at least one*. This is most useful when combined with the *[...]* symbol above, for identification of patterns with flexible gaps. The derivation rule can be rewritten in a form: $\langle q \cdots q \rangle \Rightarrow \langle q^* \rangle$.

Finally, the usefulness of empty strings in the analysis is questionable. While some practical applications can be proposed, I have chosen to disregard the empty strings and thus the Empty Concatenation derivation rule for now.

2.4.1 The restricted pattern language

According to the restrictions discussed above, the particular pattern language P_Σ we are interested in contains, besides the source alphabet Σ — this are the *atomic tokens*, the set of categories $\mathbb{C} \subseteq \mathcal{P}(\Sigma)$ — the *non-atomic tokens*, where $[...] \in \mathbb{C}$, $[...] = \Sigma$ ($[...]$ matches any token) and the repetition operation *.

The syntax of P_Σ is very similar to the syntax of $Regex(\Sigma)$:

1. $\langle a \rangle, a \in \Sigma \cup \mathbb{C}$ is in P_Σ (set of tokens)
2. If $\langle r \rangle, \langle s \rangle \in P_\Sigma$ then $\langle rs \rangle \in P_\Sigma$
3. If $\langle r \rangle \in P_\Sigma$ then $\langle r^* \rangle \in P_\Sigma$

The semantics of P_Σ is a sufficient subset of the semantics of $Regex(\Sigma)$:

1. $\langle a \rangle, a \in \Sigma$ matches the string a
2. $\langle X \rangle, [X] \in \mathbb{C}$ matches any string s where $s \in [X]$
3. $\langle r^* \rangle, r \in \Sigma \cup \mathbb{C}$ matches a string $a_1 \cdots a_n$ where $\langle r \rangle$ matches each a_i. The repetition operation applies to the single token it precedes. Also note that $x^* x^*$ is equivalent to x^* so these patterns are treated as being equal.
4. $\langle rq \rangle$ matches a string $s_1 s_2$ such that $\langle r \rangle$ matches s_1 and $\langle q \rangle$ matches s_2

Clearly, P_Σ is a simpler subset of $Regex(\Sigma)$: it has no explicit optionality (only predefined optionality via $\mathbb{C}$), no empty strings and no full-scale Kleene closure (only token repetition). The pattern generation theorem, applied to this language, consists of two deprecation rules only.

Theorem 2.4.1. ***Restricted pattern construction***

Given a string s*, every restricted pattern that matches* s *can be recursively constructed from the trivial restricted pattern* $\langle s \rangle$ *using only — and only — the following rules*

1. $\langle r \rangle \Rightarrow \langle X \rangle$*, where* $r \in \Sigma$ *and* $[X] \in \mathbb{C}, r \in [X]$ *(symbol replication)*
2. $\langle r[\cdot r] \rangle \Rightarrow r^*$*, where* $r \in \Sigma \cup \mathbb{C}$ *(repetition)*

Proof. The proof is very similar to the proof of the full theorem on the page 28. For a string of length one, there is a finite number of patterns matching it. More

specific, given a string $a, a \in \Sigma$, it is matched only and only by the patterns $\langle a \rangle$ and $\langle a^* \rangle$ and $\langle [X] \rangle$, $\langle [X]^* \rangle$ where $[X] \in \mathbb{C}, a \in X$. These are the patterns covered by the derivation rules and thus, they are reducible to the trivial pattern $\langle a \rangle$.

Now, let us assume that the theorem is valid for any string s with length less than n and consider a string $sa, a \in \Sigma$. If $\langle r \rangle$ matches sa, its form is restricted to $\langle pq \rangle$ where $\langle p \rangle$ matches $\langle s \rangle$ and $\langle q \rangle$ matches a. According to the induction assumption, $\langle p \rangle$ is reducible to $\langle s \rangle$. Because a is a single symbol, $\langle q \rangle$ is reducible to $\langle a \rangle$ and thus $\langle r \rangle$ is reducible to $\langle sa \rangle$. A degenerate case where $\langle p \rangle = \langle q \rangle = \langle r^* \rangle$ is not a problem here, because we have stated earlier that $\langle r^* r^* \rangle = \langle r^* \rangle$. □

Hence, the patterns matching s can be constructed in two steps. First, we construct all permutations by replacing each symbol in s with the corresponding group from $\mathbb{G}$. If a symbol a_i is part of n_i groups, then there are $\prod(n_i + 1)$ permutation total. Second step is the application of the repetition rule. The total number of generated patterns is therefore finite and the algorithm terminates.

2.5 The framework and its implementation

In the preceding pages I have laid out the theoretical foundation to the problem of finding patterns in corpora. Using these results, a practical solution to the problem can be devised. In following, I present a simple version of a framework for frequent patterns identification and discuss the details to its implementation.

The framework itself is comprised of a number of computer algorithms, guidelines and a particular workflow. The framework operates on a POS-tagged corpus formatted in a special way. The utterances of a corpus are considered to be strings of a formal language (completely defined by the corpus). The alphabet of this formal language are unique linguistic tokens which constitute the individual utterances. The choice of tokens is empirical; one may want them to be words, morphemes, phrases or combinations thereof.

The framework generates all regular expressions (patterns) following the rules of the restricted pattern language (2.4.1) which match at least one utterance of the corpus. The patterns which are considered infrequent (based on an independent characteristic) are then removed. A particular issue involves pattern overgeneration: the algorithm will produce sets of similar patterns, many of which will be of limited interest. For this reason, the framework includes algorithms which are designed to detect and eliminate such patterns.

The computer algorithms of the framework are implemented in R (http://www.r-project.org/) with some routines implemented in C for increased performance. The R environment is a natural choice because of its flexibility and ease of use in respect to the way the corpus data can be handled and transformed. In addition, R provides a full-fledged, high-level programming language and rich selection of statistical algorithms.

A step-by-step workflow when using the framework involves:

1. Encode the POS-tagged
    a) Choose the individual tokens
    b) Describe groups of token categories (if any)
2. Generate the patterns

   a) Execute the pattern generation routine for each utterance, collect the patterns and remove the duplicates

3. Compile the match tables[8]

   a) For each generated pattern, determine the set of utterances in the corpus it matches

4. Select frequent patterns

   a) Remove all patterns whose match frequency is below a threshold

5. Pattern filtering

6. Pattern analysis

2.5.1 String encoding

After the choice of individual tokens has been made, each token is assigned a unique integer number. This allows us to represent each utterance in the corpus as a sequence of numbers. The reason behind this is to simplify and speed up the implementation of algorithms. Formally, the corpus is a formal language L which is comprised of number sequences as strings.

The framework supports predefined groups of token categories (see 2.4). A category description simply includes the identity of all token it includes. A single token may be a member of more than one category. A special category, which is always defined is the *[...]* category, which includes all tokens. This category is used to describe gaps in patterns.

The resulting patterns are also encoded as sequences of numbers — the pattern language P_L is a strict superset of L. In addition to the linguistic tokens from the corpus, the pattern language also includes the token categories (which are also assigned a positive integer number). The numbers within a pattern may also be negated to encode repetition.

The following example illustrates the encoding process. Consider an English corpus which contains the sentences *I see a cat, I see a dog, I see a car*. We choose the tokens to be individual words. Thus, there are six tokens in total: *I, see, a, cat, dog, car* which are assigned numbers from 1 to 6 respectively. Then, *I see a cat* is

[8] A custom pattern matching routine was implemented for efficiency reason. I will not discuss the implementation details of the routine here.

encoded as $(1;2;3;4)$. In addition, we define a category of nouns which includes tokens *dog, cat, car*. This category is assigned the number 1000 and the match-any *[...]* category is assigned the number 1001. Below are some examples of how patterns which match this corpus are encoded:

(7) **Sample patterns and their encoding**

a. $\langle$*I see a [N]*$\rangle$
$\langle 1;2;3;1000\rangle$

b. $\langle$*I see [...] N*$\rangle$
$\langle 1;2;1001;1000\rangle$

c. $\langle$*I [...]**$\rangle$
$\langle 1;-1001\rangle$

d. $\langle$*[...]* a N*$\rangle$
$\langle -1001;3;1000\rangle$

2.5.2 Pattern generation

The pattern generation algorithm directly implements the derivation rules from the page 34. This algorithm is divided in two steps. First, all patterns of the same length as the input utterance are generated. If the utterance consists of n tokens $s_1 \cdots s_i$ then the sought-after patterns given by:

$$\{a_1 \cdots a_i | a_i \text{ matches } s_i\}$$

That is, each atomic token is replaced by a set all tokens which match it — this are the particular token itself and the token categories it belongs to. If the atomic token s_i belongs to k_i categories ($k_i \geq 1$, as each token belongs to [...]), then the number of generated patterns equals $\prod(k_i + 1)$.

The second step is to shrink the gaps in the patterns. Here, sequences of [...] are replaced by $[...]^*$. The algorithm in its current version does not eliminate sequences of any other tokens as I could see no practical application for such feature (as the same token seldom occurs in a sequence in a language). The output at this stage are patterns with flexible gaps, like $\langle X\,[...]^*\,Y\rangle$ (match X, then arbitrary sequence, then Y).

The disadvantage of the algorithm is that it quickly reaches computational limitations when processing large utterances. For instance, a sentence with 20 morphemes, where 10 morphemes belong to one group besides [...], results in $2^1 0 3^1 0 =$

60466176, or about 60 million patterns. The performance penalty associated with the time and space consumption required to precess such large sentences is in no comparison to the degree of information that is obtained from them. Fortunately, such sentences are very rare. Therefore, my solution — albeit a crude one — is to simply ignore such large sentences. In a preprocessing step, a number of potentially generated patterns is computed for each utterance in the corpus and the utterances which exceed a given threshold are filtered out. The actual threshold is the matter of personal taste and should be determined empirically for the particular corpus.

2.5.3 Pattern filtering

One substantial problem with automatic pattern generation is the potentially very large amount of semi-equivalent patterns[9] generated by the algorithm. Consider an example below.

(8) a. **Utterances:**
This is a dog
This is a car
This is a lamp

b. **Patterns:** (incomplete list)
⟨*This is a dog*⟩
⟨*This is a car*⟩
⟨*This is a lamp*⟩
⟨*This [...]**⟩
⟨*This [...] a [...]*⟩
⟨*This is [...]**⟩
⟨*This is a [...]*⟩

Clearly, only ⟨*This is a [...]*⟩ is interesting to us, as this is the most specific pattern which is still able to describe the whole class of the relevant utterances. The problem is recognizing such patterns and filtering out the rest. Manual (human) selection is unpractical, as it would take too much time and is error-prone. Below, I discuss two cases where automatic filtering algorithms can be devised.

[9] patterns which describe roundly the same linguistic material

2.5.3.1 Filtering of patters with flexible gaps

One class of semi-equivalent patterns includes patterns with gaps, which are equal except the repetition operator:

(9) ⟨*[...]* tell [...]*⟩
⟨*[...]* tell [...]**⟩
⟨*[...] tell [...]**⟩
⟨*[...] tell [...]*⟩

Clearly, all above patterns are in fact instances of ⟨*[...]* tell [...]*⟩*. The filtering procedure here is very simple:

$X_0 \leftarrow$ set of sets of patterns with gaps which are equal except for the repetition operator
$X \leftarrow \emptyset$
for each $a \in X_0$ **do**
  $X \leftarrow X \cup \{$ pattern in a with the highest match count $\}$
end for
return X

2.5.4 Filtering of mutually ambiguous patterns

Mutually ambiguous patterns are the patterns which match exactly the same sequences[10]. An example of such patterns (given a suitable corpus) are ⟨*This is [...]**⟩ and ⟨*This is a [...]*⟩ from the Example (8). As their definition suggest, mutually ambiguous patterns can be identified trivially.

Given a set of mutually ambiguous patterns, the most interesting pattern is the one which is most specific, that is, the patterns which retains the maximal amount of fixed linguistic material. To find this pattern, we simply count the number of non-gap elements in the patterns and select the pattern with the largest value. There is only one such pattern, which is easy to prove.

Let r_1 and r_2 be different mutually ambiguous patterns which both are most specific, that is, they both have n non-gap elements and there is no other pattern which is mutually ambiguous to r_1 and r_2 and has more than n non-gap elements.

[10] strictly spoken, patterns with flexible gaps also are mutually ambiguous patterns. However, as a special case, they require a separate treatment.

Now, as r_1 and r_2 are not identical, they must have at least one distinct element. That is, r_1 must have a non-gap element where r_2 has a gap and r_2 must have a non-gap element where r_1 has a gap. But this means, we can construct a pattern r' which has non-gap elements in both this positions and which is able to match anything that r_1 or r_2 can jointly match. The pattern r' has $n+1$ non-gap elements and it is mutually ambiguous to r_1 and r_2. Thus, no such r_1 and r_2 can exist. □

The following algorithm can therefore be safely used to eliminate the mutually ambiguous patterns.

$X_0 \leftarrow$ set of sets of mutually ambiguous patterns
$X \leftarrow \emptyset$
for each $a \in X_0$ **do**
  $w_i \leftarrow$ number of non-gap elements in $x_i \in a$
  $X \leftarrow X \cup \{$ pattern in a with the highest $w \}$
end for
return X

2.5.5 Intermediate summary

In this chapter I have described a framework for identifying frequent morphosyntactic patterns in POS-tagged language corpora. The core idea of the framework is to use regular expressions in order to represent the patterns. The framework relies on the theoretical results from the Theorem 2.2.1 to generate all relevant regular expressions which describe the sentences in the corpus. In the subsequent step the patterns are filtered according to their frequency.

Here, I want to briefly summarize the benefits of my framework over the previous approaches. First, as already discussed, the methods used in previous studies were hand-tailored for the immediate purpose. They were only able to detect what they were expected to detect and in addition, difficult to extend.

My framework, taking the very general approach, has none of that shortcomings. The pattern generation is based on a strictly mathematical proof which ensures that no patterns are undetected. Of course, one may accuse me of being a bit of a hypocrite here, as the constraints I have discussed in 2.4 can potentially lead to exclusion of relevant patterns. However, the constraints have been carefully designed in order to exclude such situations.

The framework is able to capture a large number of pattern types. Below I list some possible patterns just to give the reader an impression. The list is not exhaus-

tive!

Uninterrupted sequences

⟨*[...]* super -cali -fragil -ist -ic -exp -iali -doc -ious [...]**⟩
⟨*super -cali -fragil -ist -ic -exp -iali -doc -ious [...]**⟩
⟨*[...]* super -cali -fragil -ist -ic -exp -iali -doc -ious*⟩

Sequences with gaps

⟨*I [...]* cats*⟩
⟨*I and [...]* like [...]**⟩
⟨*[...]* and [...]* show [...]**⟩

Sequences with categories

⟨*I see [N]**⟨
⟨*[...]* [V] the [N] [...]**⟨
⟨*[Pro] [V] [Det] [Adj] [N]*⟨

Because of its flexibility, the framework or its components can be utilized in a large number of tasks. It is equally suitable to the analysis of the child language as to the analysis of child-directed language. There is a number of additional interesting possibilities. I discuss some of them in the conclusion part of the thesis.

Part III

A case study: child-surrounding speech in Chintang

The patterns emerge...

3.1 Chintang and its people

Chintang is an Eastern Kiranti (subfamily of the Sino-Tibetan language family) language, which is spoken by a small community in the Chintang VDC[11] in Dhankuta District, Koshi Zone, Eastern Nepal. The current number of speakers is estimated to be about 5000. The Chintang people live in villages in a rural, hilly area, with their primary occupation being farming.

The Chintang language is a highly endangered language, and is currently experiencing language shift to the dominating neighboring languages Bantawa and Nepali. There are virtually no monolingual Chintang speakers, as they are usually proficient in Bantawa or/and Nepali — the official lingua franca of Nepal. A significant part of current Chintang lexicon is borrowing from Nepali. Still, Chintang is the language that is spoken at home and the first language that the children acquire.

A detailed documentation of Chintang started only recently, in 2004, with the *Chintang and Puma Documentation Project* (further CPDP) — a DoBeS project carried out by the linguistic departments at the University of Leipzig, Germany and Tribhuvan University, Kathmandu. The Project is funded by the Volkswagen foundation (Grant № II/79 092, 2004-2008 PI Balthasar Bickel). Thus, Chintang is still an unresearched area in many regards. There are not many publications on Chintang either (Bickel et al., 2005a,b, 2007; Stoll et al., 2008, e.g.).

A significant part of the collected data is the POS-annotated language corpus, with a large part of the corpus devoted specially to language acquisition.

3.1.1 Chintang from a typological perspective

Chintag is a highly polysynthetic language, especially in regards to the verbal morphology. The basic alignment is ergative. The verbs agree with both subject and object in person and number. There is a large number of inflectional categories, including tense, aspect, polarity and mood. In addition, Chintang has compound verbs and verb incorporation. All verbs act alike, and only a few number of irregular verbs are present in the language. Thus, there is basically only one — albeit very extensive — paradigm for verbal inflection. As Chintang is subject to a (so

[11] *Village Development Committee*, a municipal administrative unit

far unique) grammatical curiosity — free prefix permutation (Bickel et al., 2007), the observed number of inflected forms is even higher and can be as much as 1359 (Stoll et al., 2008, c.f.). In comparison, the noun morphology is rather "simple" — there are two numbers (singlular vs. non-singular) and eleven cases, obligatory marked in noun inflection. There is also a large number of distinct discourse particles.

The word order in Chintang is verb-final, usually SOV. However, the word order is rather relaxed and allows variations. The wh-particles are not required to be fronted and may stay in situ. Chintang is also a subject to massive argument drop, resulting in the fact that a Chintang sentence often consists of the inflected verb form only.

Chintang is a particularly interesting case for the cross-typological research on lexical restrictiveness as first attempted by Stoll et al. (2009). Typologically, Chintang differs drastically from the European languages which were considered by previous studies (Stoll et al., 2009; Cameron-Faulkner et al., 2003). The massive number of morphological possibilities paired with the argument drop suggest that lexical restrictiveness (like found in English, Russian and German) in Chintang is unlikely.

3.2 The corpus

The data I use here is the CPDP Chintang language acquisition corpus. The corpus contains longitudinal recordings of four Chintang preschool children, with first two children staring at the age of 2 and the last two children starting at the age of 3. All children reside in large houses together with their family, have at least three siblings and come from different households. Chintang was the preferred language used by children and adults in their interaction.

The recordings took place in a natural environment: a camera and a microphone were placed in the location where the children played, usually near the house. A Nepalese research assistant, together with local native-speaking assistants, took care of the technical equipment, sometimes interacting with children. No particular situation was enforced, rather, children were recorded during their usual interaction with other people. Such recording style resulted in a large number of contexts captured by the corpus, often including various conversation between other children and / or adults. Most importantly, not the exclusive interactions with the target child are captured, but rather, all interaction which takes place during the recording. This makes the corpus a particularly natural approximation to what the child hears in its everyday life, as Chintang children (unlike many children in modern society) do not stay at home, interacting with their parents or a selected number of caretakers exclusively, but instead, roam around the village in groups, playing and occasionally interacting with other village people. Thus, the natural description of the corpus involves the notion of *child-surrounding* rather then child-directed speech, and it is child-surrounding speech I will be concerned with in this study[12].

The recordings were taken in regular intervals and covered eighteen months in total. Each child was recorded for four hours per month. These recordings took place during a single week, and were distributed over as many individual recording sessions as necessary. Thus, a single month worth of recording (*round*) consisted of relatively local data chunks (*sessions*), separated by several days only. At the moment of writing of this thesis, a substantial portion of sessions was yet to be tagged

[12]In theory, it is possible to extract only child-directed interactions from the corpus. But such process would basically involve retagging of the video and audio data — an undertaking which was clearly out of my practical possibilities

and glossed. Thus, I could use only a portion of the data, which included the first fourteen rounds per child (effective fourteen months worth of recording), with possible gaps in between. The quantitative and qualitative evaluation of the corpus is provided in the next section.

3.3 Pattern identification

3.3.1 Corpus preparation

In the first step, the corpus was transformed in a representation suitable for further analysis. As I was interested in child-surrounding speech only, all sentences uttered by target children were removed from the corpus. No distinction was made between particular speakers, rather, all child-surrounding speech was treated as if uttered by a single speaker. This is legitimate, because, as already noted, the surroundings of the children are very dynamic and children are involved in interactions with a large number of adults. Duplicate utterances were not removed, as doing so would alter the natural input (various control mechanisms were employed to ensure that duplicate utterances did not lead to "false" frequent patterns — see below in text).

The remaining data was encoded in a manner described in 2.5.1. The individual tokens were morphemes and not words. This seemed as the more appropriate choice, given the rich morphology of Chintang. Overall, there were 5604 distinct atomic tokens (morphemes). For the sake of simplicity, I did not use any token categories (see 33) besides the obligatory gap category *[...]*. After the overly long utterances were removed (see page 38, the threshold was set to 65536 patterns), the final corpus contained 173708 tokens in 40099 utterances.

For the purpose of the analysis, the corpus was divided into continuous chunks of data. The reasoning behind such partitioning was the following consideration. It is important to ensure that the frequent patterns which one discovers in the corpus are uniformly distributed over the longitudinal recordings — that is, the high occurrence frequency must not be a local phenomena only. Patterns which are just locally frequent — for instance, within a single conversation — are of no interest to the study, as they are not a long-term frequent component of children's input. Cameron-Faulkner et al. (2003) propose a criterion for pattern frequency. In that study, a pattern (*frame*) was considered locally frequent if it occurred at least 4 times in 1400 utterances. A pattern was considered globally frequent if it was locally frequent in at least half of the subcorpora (Cameron-Faulkner et al. (2003) refer to them as *core* frames) that is, it had to be frequent both within the subcorpora and between the subcorpora. The same criteria was also adopted by Stoll et al. (2009).

As the choice of criteria is purely empirical, I follow their example with my study. The only difference was that the previous studies used subcorpora which contained exactly (or around) 1400 utterances. In my study, data chunks had variable sizes, hence, the frequency criteria I used were relative to the chunk size and equaled $\frac{4}{14}\%$. Thus, a pattern was locally frequent in a chunk if it accounted at least for about 0.28% of the chunk utterances.

Table 3.1: Data chunks (distribution and size)

Chunk	I		II		III		—	IV		V		—	VI	
Month of recording	1	2	3	4	5	6	7	8	9	10	11	12	13	14

	Chunks					
	I	II	III	IV	V	VI
Child 1	1223	2597	1721	2013	1303	1912
Child 2	520	1748	906	2820	1422	1332
Child 3	572	961	1029	1695	2426	1904
Child 4	1064	2054	1351	2059	2743	2724

The chunk partitioning was not arbitrary, but followed clear guidelines. First, the corpus was split into four groups, in respect to the target children (further in text I will refer to this groups as child subcorpora). Then, each child's subcorpus was split into a number of parts (the actual chunks), such that each chunk contained about two months (more precisely, two weeks in consecutive months) worth of recording. Some chunks contained only one month worth of recoding data, due to the gaps in the available data. This particular size of chunks was chosen as it resulted in the most balanced distribution. The distribution ensured that similarly numbered chunks from different child subcorpora corresponded to roughly the same months of the recording. The Table 3.3.1 shows the correspondence of chunk to recording months and the final counts of the utterances per chunk.

Note that about the half of the chunks in fact contained less then 1400 utterances. This can potentially pose a problem for frequent pattern identification (given the threshold of 4 in 1400), due to the reduced accuracy. In particular, the first chunks from the second and third child subcorpus, with 520 and 572 utterances respectively, are especially pathological cases, as $\frac{2}{572} \cdot 100\% = 0.34\%$. This means that every generated pattern which matches at least two utterances in these

chunks will be considered frequent[13]. Other chunks may too overgenerate , with results becoming more accurate as the chunk approaches the magic 1400 length.

However, this issue is not as crucial as it may appear at first. Consider the following. There are two problems connected with the small-sized chunks. First, they may overgenerate (where less-frequent patterns are falsely reported as more frequent). Secondly, they may undergenerate (potentially high-frequent patterns are not encountered due to reduced sampling resolution). The first problem concerns only chunks with the size below 1400. The second problem potentially applies to *any* chunk, as even large amounts of data are not guaranteed to contain all high-frequent patterns of the language. If the resolution of at least 1400 utterances is a priory assumed to be sufficient, all chunks of comparable size should be sufficient as well.

Now, only 2 chunks out of 24 are pathological, 14 contain more then 1400 utterances and out of 8 chunks left, 4 have more then 1200 utterances. Thus, 18 of 24 chunks should have sufficient resolution for accurate pattern generation, and 10 of 24 chunks potentially overgenerate. The later fact is rendered immaterial by the global frequency criterion: even if an overgenerated pattern is present in all the 10 chunks, it still will not be registered — as at least 12 chunks are required to consider a pattern frequent. In conclusion, the undersized data chunks in our distribution are very unlikely to distort the results of the study.

3.3.2 Pattern generation and comparative chunk evaluation

For each utterance in the chunk, the algorithm generated a set of patterns which match this utterance. All such patterns were collected in a single list, with the duplicates removed. Patterns with the relative frequency less then $\frac{4}{14}\%$ within the chunk were discarded. The final result was the set of frequent morphosyntactic patterns per data chunk (Table 3.3.2).

The number of generated patterns for the majority of the chunks was roughly between 800 and 1300 patterns. The numbers are not normally distributed (Wilk-Shapiro test (Royston, 1995, test) for normality reports a high significant p-value of $7 \cdot 10^{-9}$). Out of 24 chunks, three are clear outliers: the first chunks of the second, third and fourth child subcorpus, respectively. The first two are the pathologically under-sized chunks, which explains the high number of generated patterns.

[13] This will apply for any chunk size under $2 \cdot \frac{4}{1400} = 700$ utterances

Table 3.2: Frequent patterns per chunk (unfiltered)

	Chunks					
	I	II	III	IV	V	VI
Child 1	1946	1049	1130	948	1245	887
Child 2	6018	1250	2040	861	939	2909
Child 3	5503	177	1255	1385	947	874
Child 4	21463	989	1443	900	881	850

The exceptionally high number of patterns associated with the first chink of the fourth child subcorpus goes unexplained at first.

However, the numbers from the Table 3.3.2 are potentially misleading, as they contain large number of semi-equivalent patterns (as discussed on page 39). I have applied automatic filtering mechanisms described in the last chapter to eliminate such patterns, thus reducing the pattern set to the most specific ones. This resulted in a new distribution which is shown in the Table 3.3.2.

Table 3.3: Frequent patterns per chunk (filtered)

	Chunks					
	I	II	III	IV	V	VI
Child 1	707	601	639	547	721	512
Child 2	435	650	656	425	536	657
Child 3	646	549	681	699	544	502
Child 4	700	537	781	534	580	510

As one can see, the filtering changes the pattern counts quite significantly. First, the outliers are completely eliminated[14]. Second, the result is much more dense then with unfiltered patterns: Levene's test (Levene, 1960) of equal variance produces p-value of 0.0074[15]. Third, the numbers of filtered patterns form a normal distribution: the Whilk-Shapiro test results in a p-value of 0.5.

An important issue which I want to discuss here concerns the independence of the data chunks. There is a distinct possibility that the distribution of the patterns within the chunks are affected by additional variables, thus resulting in a particular

[14]Most notably, the initial number of 21463 patterns for the chunk 4_I was reduced to 962 patterns only. The high initial number of patterns could be explained by some peculiarities in the data chunk: for instance, duplicates of long utterances could produce a large number of similar frequent patterns. This particular example shows the significance of filtering — manual reduction of over twenty thousand patterns would not be manageable.

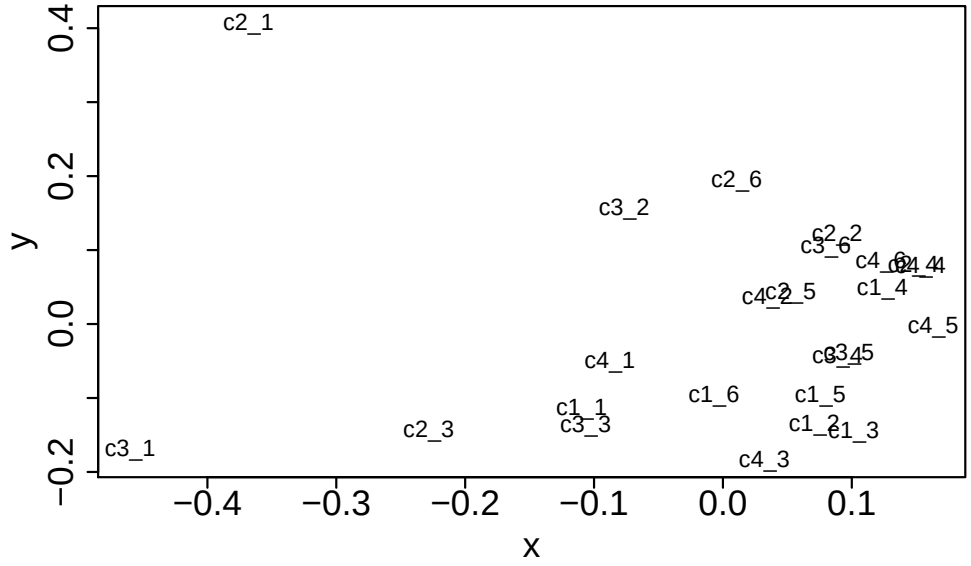
[15]Levene's test was chosen as it does not require normality of the data

chunk ordering. For instance, the pattern distribution could be child dependent — then, the chunks of the same child subcorpus will show similarities between each other, i.e. they will form a cluster. The age of the children could be another factor: in this case, similarity between the chunks from different child corpora will be observed, depending on the age of children at the point of time of the recordings covered by the particular chunks.

If the chunks are not entirely independent, the whole idea of pattern analysis is compromised. Chunk dependency effectively reduces the amount of available data — if data sources are related, they cannot be considered as independent sources of evidence.

In order to visualize the similarities between the individual data chunks, I have employed a multidimensional scaling (MDS) technique. The MDS attempts to find a point distribution in a n-dimensional Euclidian space, such that the distances between the points approximate the distance between the original data elements. When performed in the 2-dimensional space, MDS produces a convenient "flat" plot which graphically represent the similarity of the data chunks. The closer chunks are on the plot, the more similar they are. Visual inspection of the plot allows for easy identification of chunk clusters.

Figure 3.2: MDS plot of chunk similarity



A very simple distance measure was used to estimate the degree of similarity between the individual chunks. Chunks were regarded as similar if they shared a large number of identical patterns. The distance was computed as $1 - \frac{|A \cap B|}{|A \cup B|}$, where A and B are the sets of filtered frequent patterns in two different chunks. A distance matrix, containing pairwise distances between distinct chunks, was then computed. The MDS was applied to the matrix, producing the Figure 3.3.2.

The MDS plot reveals no apparent clustering at all. Notably, the pathological chunks c2_I and c3_I clearly show on the plot as the chunks with the greatest distance to the rest of the group.

Thus, the preliminary evaluation leads to the conclusion that the data chunks are in fact independent of each other. This means that Chintang children receive input independent of their identity and their age (at least in the situations covered by the CPDP recordings).

3.4 Pattern analysis

3.4.1 Distribution

Overall, 332 distinct patterns with frequency at least 4 in 1400 utterances were discovered in at least 12 of 24 data chunks, or 50% of chunks. This are the globally frequent patterns (c.f. to *core frames*). These 332 patterns together matched 31788 of 40099 utterances, or roundly 79% of the whole corpus. It is important to note that the patterns are not mutually exclusive, that is, two different patterns can match the same utterance. The Table 3.4 shows a general overview of the globally frequent patterns in respect to their frequencies within and between the chunks.

The table shows that the majority (264) of patterns each matched less then 500 utterances in the corpus (or about 1.3% of the corpus). Only 14 of such "lightweight" patterns were present in all of 24 chunks, a majority of them were only encountered in 12 to 18 chunks. On contrary, the "heavyweight" patterns, with more then 500 matches were more frequently found between the chunk; with patterns which matched more then 1000 utterances found almost exclusively in every chunk. Such distribution can be trivially explained: the lower the pattern frequency, the lower the chance that it will be encountered in a chunk.

Table 3.4: Number of globally frequent patterns within the data chunks and their distribution

	Occurrence in the chunks			
Matches	12 to 17	18 to 23	24	**Total**
Less then 200	111	14	0	125
200 to 499	40	85	14	139
500 to 999	4	20	20	44
1000 or more	1	2	21	24
Total	156	121	55	332

3.4.2 Pattern types

I have designed a simple two-dimensional dichotomy in order to classify the patterns. One dimension was the position of the (non-gap) tokens within the pattern (utterance initial, utterance final, middle or both initial/final). Another dimen-

sion was the occurrence of gaps between two tokens (gap or no gap)[16]. The below tabular illustrates the dichotomy:

	Position			
Has gap	**Initial**	**Final**	**Middle**	**Initial/Final**
Yes	⟨*X [...] Y [...]*⟩	⟨*[...] X [...] Y*⟩	⟨*[...] X [...] Y [...]*⟩	⟨*X [...] Y*⟩
No	⟨*X [...]*⟩	⟨ *[...] Y*⟩	⟨*[...] X [...]*⟩	⟨*X*⟩

The Table 3.5 shows the breakdown of the found patterns according to the dichotomy. Following can be observed. There are more patterns without gaps than with gaps (271 to 61). Patterns with neither initial nor final position of lexical elements are the majority. Also, there are more utterance-final patterns then utterance-initial (86 to 50). A Monte-Carlo significance test (Hope, 1968) with 200000 replicants produces a p-value of 0.1, thus indicating that there is only slight chance for any correlation between the position of the lexical material in the pattern and the presence of gaps.

Table 3.5: Pattern classification and counts

	Position				
Has gap	**Initial**	**Final**	**Middle**	**Initial/Final**	**Total**
Yes	7	23	30	1	61
No	43	63	154	11	271
Total	50	86	184	12	332

The Table 3.6 shows the number of non-gap tokens (morphemes) within the patterns and the corresponding pattern count. As we can see, patterns in Chintang are rather "short", with the overwhelming majority of pattern featuring only one or two fixed morphemes.

The exhaustive list of all discovered patterns and the used abbreviations are in the Appendix A. The utterance-medial patterns constitute the clear majority (with one hundred eighty four patterns) of total frequent patterns. Of these,

[16]Note that the initial/final position implicates either a gap between two elements or a totally rigid pattern (pattern which matches one utterance only). As the matter of fact, all but one such pattern I encountered in the corpus consisted of a single-word utterance like "yes" or "what". As mono-morphemic utterances are not particularly interesting, I will not discuss such patterns further

Table 3.6: Number of lexical items within the patterns

Number of morphemes	1	2	3	4
Patterns	200	113	18	1

106 are not particularly interesting as they consist of one token only. These patterns simply show frequent morphemes (like tense markers). The remaining 78 patterns were frequent collocations of various morphemes. This could be frequent co-occurrences of verbal markers like in ⟨*[...]* -a|IMP -⊠|EMPH [...]**⟩ (which suggests that imperatives are further stressed with the help of emphatic marker — see also page 57). Another example are frequent verb form like ⟨*[...]* khat|go -a|PST [...]**⟩. As the detailed analysis of these patterns deserves its own study, I will not discuss the utterance-medial patterns more closely but instead focus my attention on the utterance-initial and utterance-final patterns.

3.4.3 Utterance-initial core patterns

Overall, 50 utterance-initial core patterns were reported which account for 13837 utterances, or 34.5% of the total corpus. Of this 50 pattern, 43 had no gaps. This patterns were very short, containing just one morpheme, the only exception being two initial two-morpheme collocations ⟨*akka|1s -ko|GEN [...]**⟩ and ⟨*yo|DEM.ACROSS -ni|DIR [...]**⟩. The majority of initial morphemes consisted of wh-particles, demonstrative pronouns and pronouns, as illustrated by the loose classification in the Table 3.7. Morphemes which I here classified as *Other* included negation, mood particles[17] and temporal determiners like "today" and "now". The last group, *Addressing* contained all morphemes used to address another person directly. This included names, imperative verb forms and similar.

Table 3.7: Utterance-initial frequent morphemes

Wh-particles	**Pronouns**	**Demonstratives**	**Addressing**	**Other**
14%	21%	23%	14%	28%

The remaining 7 utterance-initial patterns with gaps are frequent collocations of a frequent utterance-initial morpheme and a verb marker, they contain two mor-

[17] Chintang has a relatively large number of emphatic interjections

phemes with a gap in between. For instance, one such collocation is ⟨*akka|1s [...]* -u|3P [...]**⟩. The *-u* marks object agreement, and the pattern thus describes utterances where the first-person actor is acting on a third-person object.

3.4.4 Utterance-final core patterns

In total, 86 patterns were utterance-final, with 63 patterns matching a continuous sequence of morphemes (no gap). The majority of this patterns (42) matched only one single morpheme at the end of the utterance (Table 3.8). This patterns account for 20582 utterances, or 51% of the whole corpus. Most of frequent utterance-final morphemes were either discourse particles, verb suffixes (tense / agreement markers) or nominal suffixes (case, number markers). Morphemes marked as *Other* in the table included, among others, a wh-particle and a number of adverbs.

Table 3.8: Utterance-final frequent morphemes

Discourse particles	**Verb morphology**	**Noun morphology**	**Demonstratives**	**Other**
28%	26%	14%	5%	26%

The remaining utterance-final patterns — containing more then one fixed morpheme — describe frequent collocations of frequent final morphemes with some additional markers in the utterance.

The discourse particle *na* One particularly interesting collocation involves the utterance-final particle *na*. It is one of Chintang discourse particles and the most frequent utterance final morpheme (found in this position in 2501 utterances, or 6.3% of the corpus). Inspection of respective patterns reveals that this particle is strongly associated with the imperatives: 1636 utterances which ended on *na* also contained an imperative. Another frequent collocation involves the verb *lut* (to tell) — 403 utterances which ended on *na* contained a form of *lut*.

3.5 Summary and outlook

One important result from (Stoll et al., 2009) is that English, German and Russian are similar in regards to the lexical repetitiveness at the beginning of the sentence: just a limited variation of sentence-initial frames accounted for relatively high amount of utterances in all three languages. Hereby, differences based on the typological properties were observed. For instance, English shows the highest number of frequent utterance-initial patterns (compared to German and Russian) while Russian has a smaller amount of such patterns than English and German, in addition, such patterns in Russian tend to be shorter. This can be easily explained, as English is a language with rigid word order and poor inflectional morphology, where helper devices such as copula, articles and auxiliaries (which usually occur close to the beginning of the utterance) are obligatory. The combination of these factors produces a higher number of (potentially longer) frequent initial sequences. In Russian, the situation is quite different, as this language has relaxed word order, rich inflectional morphology and less auxiliary use. For instance, instead of saying *This is a X* a Russian speaker uses *Eto X* (lit. *This X*), thus effectively saving two words in the pattern. In short, English largely relies on syntax to encode constructions where Russian more eagerly relies on morphological markers. This results in the frame size differences. German, as a language which typological properties lie somewhere in between, has more utterance-initial variation than Russian but less than English. The frequent utterance-initial patterns accounted for about 64% of produced utterances in English and German, and 57% produced utterances in Russian.

However, despite this differences, the similarity between the three languages was striking. All of them showed substantial amount of repetitiveness and thus, predictability, at the prominent utterance initial position — as a relatively small number of words accounted for more than a half of all produced sentences. Another important similarity is the content of frequent utterance-initial words: all three languages followed a similar trend here. Most frequent forms included pronouns, demonstratives, imperatives and wh-particles — even if the language grammar does not explicitly request obligatory fronting of this elements.

Clearly, Chintang, when compared to the three European languages, is more

typologically "extreme" than Russian, as the word order in Chintang is even more relaxed and the inflectional morphology particularly rich. So how does the analysis of frequent patterns in Chintang compare to the results of (Stoll et al., 2009; Cameron-Faulkner et al., 2003)?

Before I discuss this question in more detail, a disclaimer has to be issued. Direct comparison of results between these two studies is a dodgy undertaking, as strictly spoken, the studies analyze different types of speech. While in (Stoll et al., 2009; Cameron-Faulkner et al., 2003) only child-directed speech was considered, my study was performed on rather "everyday" Chintang, which included both child-directed speech and conversations between adults. It should be generally expected that child-directed speech shows overall more repetitiveness and less complexity then child-surrounding speech. However, as I don't have the data on the Chintang child-directed speech specifically, for the comparison I have to assume that the differences between the styles is insubstantial. While such assumption is hardly plausible, it at least allows us to describe roughly similar trends — if any should exist — in both studies.

Even as a rough estimate, it is not difficult to see that the results I have described on the last few pages clearly follow the trend established in (Stoll et al., 2009; Cameron-Faulkner et al., 2003). First, the Chintang shows a very similar degree of lexical predictability in the prominent sentence positions. Like in Russian, frequent utterance-initial (and final) sequences are very short (usually one morpheme / word). Undoubtedly, the reason here is that Chintang grammar allows more variation than English (or German) grammar. The frequent utterance-initial morphemes only account for 34% of the corpus, which is (while still significant) much less than the corresponding figures for English / German / Russian. However, Chintang is characterized by a massive drop of verbal arguments — which could explain the low counts[18]. Even though, lower predictability at the beginning of the sentence in Chintang is offset by the high predictability at the end of the sentence. Here, 51% of the corpus is accounted for by a small set of frequent utterance-final morphemes. Together, around 64% of sentences in the corpus either begin with one of 41 initial morphemes or end in one of 42 final morphemes. Also, the types of frequent utterance-initial morphemes in Chintang child-surrounding speech is surprisingly similar to that found for English / German / Russian child-directed speech, as the majority of such morphemes are pronouns,

[18]It would be interesting to compare this number to one from Chintang child-directed speech.

demonstratives, imperatives and wh-particles.

In summary, despite the large typological distance between Chintang and the languages of Europa, all these languages show similar trends in regards to lexical restrictiveness in the prominent positions of the utterance. This leads to the suggestion that languages with quite different grammar actually have more similarity "in daily use" than some may assume.

Clearly, the analysis I presented here is very superficial. A more detailed analysis of the complex patterns, in particular the utterance-final and the utterance-medial patterns would be a logical next step. Such analysis should also include category groups like [*N*] and [*V*] to capture the distributional contexts of the respective morpheme frames more closely. It is possible that the frequent morphological patterns encountered in Chintang (and other languages with rich morphology) play a similar role in language acquisition as the syntactical (lexically restricted) patterns in languages with poor morphology, like English. This is clearly a question worth pursuing.

4

Conclusions and outlook

As I have stated in the introduction, there were two primary goals I have set for my thesis. The first — and my main — goal was to design and implement a practical framework for identification of frequent patterns in language corpora. My second goal was to put this framework to test, by making a sketchy analysis of frequent patterns found in Chintang, and to compare the results of this analysis to the results of previous studies carried out on typologically different languages. I am delighted to say that I have attained both of these goals. In following I want to briefly discuss the results and conclusions I have reached.

4.1 Pattern identification: outlook

In the final part of my thesis I have shown that the pattern identification framework which I have developed holds its own in a practical test. It was able to identify a large number of regularities in the language which previous hand-tailored methods couldn't see. Due to the general approach I adopted in designing the framework, it is particularly easy to use and tune.

My framework provides a number of crucial benefits. First, it operates on a list of abstract token sequences. No particular nature of the token is presupposed or required. In particular, it means that the framework can operate on any kind of language — the only requirement is that an adequate utterance partitioning is provided (i.e. division in words, morphemes). No knowledge of grammatical rules is required by the regular expression generation algorithm to function correctly. This makes my framework a tool which is very suitable for explorative studies in weakly researched languages.

Second benefit is that the framework is built on a mathematical theory (Theorem 2.2.1). There is a formal description of how it operates and a formal proof of why its results are correct. While one is forced to adopt a number of constrains in

regards to pattern generation (see 2.4), I have argued that the current set of constrains meets the interests of the linguistic research very well. Also, the constrains can be easily improved should such need arise.

Third benefit is that the framework is able to detect a wide range of pattern types. The design of the framework makes sure that no hidden regularity can escape. The utilization of constraints, category groups, variable matching (repetition), flexible gaps in the combination with pattern filtering (2.5.3) provides great descriptive power while retaining readability and high level of pattern relevance. Furthermore, the parts of the framework can be fine-tuned to enhance its performance in particular areas (e.g. pattern filtering).

The Chintang case study (see previous chapter) only slightly tapped the tip of the iceberg. I restricted myself to analyzing only very simple cases where the patterns matched lexical sequences in prominent positions — i.e. in the beginning and in the end — of the utterance. I have not used the advanced feature of the framework: category groups. This feature would allow to see distributional patterns around language categories, possibly providing insight on how they may be acquired. Even though, the framework has detected way more patterns that I was able to analyze. For instance, the analysis of the complex utterance-medial patterns which involve verbal morphology has to be left to the future studies.

My framework opens a number of interesting possibilities. Because the individual tokens are abstract, one can easily incorporate additional tokens into the utterances, which are not words or morphemes. For instance, such tokens could encode pragmatic and prosodic markers. Also, my framework could show some potential in areas other then corpus analysis. I can imagine it to be beneficial in the study of regularities in phonology (here a token would represent a phoneme). Also, it may prove useful in semi-automatic language exploration as the patterns may provide some insight about the language grammar (for instance, Chintang results clearly showed the agreement and case marking patterns).

Clearly, there are many ways in which the framework can be improved. In its current form it is an initial effort, hardly beyond a proof of concept. Further work would involve, in particular: *a)* refinement of pattern generation constrains to make it even more clear that no interesting pattern remains unaccounted for, *b)* refinement of the pattern filtering algorithms and *c)* overall improvements which aim to reduce the entropy of the located patterns.

4.2 The hidden structure in the language

One central result of this thesis is that Chintang, despite its large typological distance to the languages of Europe, shows similar traits when the frequent morphosyntactic patterns are considered. This suggests that there is more similarity to languages than one may see at first and clearly more than predicted by a "pure" grammar.

Where does this similarity comes from? First, despite the different languages all humans live and act in the same world with the same physical realities. Given the facts that the main reason for the very existence of language is the urge to communicate in this world and that the language is constructed economically to fulfill this function, this statistical patterns are partly simply because humans often are in the same situation. The basic hypothesis here is that the fragments of the language, which are required in frequent conversational situations organize themselves in an economical fashion. Of course, "economy" is a highly speculative notion which is hard to define formally and which has been used differently in different contexts. To discuss this notion in detail is clearly beyond the scope of my thesis. Rather, I use it in an intuitive fashion, as a principe which forces the usage of minimal means to obtain maximal effort. For instance, (virtually) every language has pronouns. The reason for this is that pronouns code the most frequent referents in the discourse. Thus, the maximal effort (reference to a broad, but clearly-defined class of entities) is obtained by minimal means (a small set of encoding tokens). It is at least possible to imagine that similar reasoning applies to other aspects of language as well. This results in certain regularities. Across the languages, this regularities may be different in *form*, but they stay similar in *spirit*, which is reflected by the statistics found in the live-speech corpora.

If this reasoning is correct, then the usefulness of such statistical regularities for an empiricist account becomes irreplaceable. If frequent patterns in the language mirror the frequent conversational needs, the task of the language learning child is simplified significantly. It would be able to identify "ìmportant" aspects quickly and starting from there, learn the language, step by step.

Under this perspective, it becomes evident why many linguists that work with formalized grammar turn to language nativism. Given the complexity and diversity found in the languages, such linguists are confronted with the vast number of abstract forms. From their point of view, it is clear that a task of learning the lan-

guage without any aid is an undertaking which is destined to be a disaster. The assumption that such aid exists in form of genetical endowment provides the only sound escape from this dilemma. However, the study of forms does not describe their daily use and a formalized grammar does not describe stochastic regularities in the language. If one detaches from the study of pure forms and in addition considers the function and the use, the language emerges as an organic, highly organized system and no language nativism may be required to motivate it's learnability per se.

Appendix A

Globally frequent pattern list

Each pattern is written as list of morphemes written as *phonological form* | *GLOSS*. The dash indicates prefix or suffix. The symbol *[...]** represents the variable gaps in the pattern. The start and end of the pattern correspond to the start and the end of the utterance. The glossing followed the conventions established in Leipzig glossing rules (http://www.eva.mpg.de/lingua/resources/glossing-rules.php)

Table A.1: Utterance-initial patterns

ba|DEM.PROX [...]* (1186)	akka|1s [...]* (959)
yo|DEM.ACROSS [...]* (658)	a-|2 [...]* (578)
a-|1sPOSS [...]* (531)	them|what [...]* (523)
hana|2s [...]* (523)	lo|yes [...]* (504)
i-|2sPOSS [...]* (493)	nunu|baby [...]* (487)
huĩ|DEM [...]* (435)	bai?|DEM.PROX [...]* (375)
la|EXCLA [...]* (371)	huŋgo|DEM [...]* (371)
u-|3sPOSS [...]* (371)	khat|go [...]* (365)
hokke|where [...]* (358)	mai-|NEG [...]* (333)
abo|now [...]* (298)	mo|DEM.down [...]* (295)
sa|who [...]* (284)	khoi|EXCLA [...]* (278)
lo|ok [...]* (265)	akka|1s [...]* -u|3P [...]* (263)

aŋ|what [...]* (254)
mo|DEM.DOWN [...]* (204)
akka|1s -ko|GEN [...]* (182)
na-|3>2 [...]* (179)
kina|SEQ [...]* (172)
thapt|bring.across [...]* (168)
yo|DEM.ACROSS -ni|DIR [...]* (162)
ci|eat [...]* (157)
bago|DEM [...]* (149)
yo|DEM.ACROSS [...]* -a|IMP [...]* (142)
ba|DEM.PROX [...]* -a|IMP [...]* (124)
la|ATTN [...]* (121)
e|INTERJ [...]* (111)
ne|take [...]* (204)
akka|EXCLA [...]* (200)
aiya|EXCLA [...]* (181)
to|DEM.UP [...]* (175)
kanchi|youngest.one.fem [...]* (172)
khoi|where [...]* (167)
akka|1s [...]* -ŋa|1sS/P [...]* (158)
ba|DEM.PROX [...]* -u|3P [...]* (153)
theke|why [...]* (142)
akka|1s [...]* -kV|NPST [...]* (126)
ã|yes [...]* (121)
paĩ|today [...]* (119)
akka|1s [...]* -ŋ|1sA [...]* (109)

Table A.2: Utterance-final patterns

[...]* na|PTCL (2501)
[...]* -a|IMP (1995)
[...]* -?|EMPH (1374)
[...]* ni|PTCL (860)
[...]* kha|PTCL (708)
[...]* mo|PTCL [...]* na|PTCL (641)
[...]* -a|IMP -?|EMPH (613)
[...]* aŋ|PTCL (572)
[...]* -e|PST -?|EMPH (520)
[...]* -kV|NPST (470)
[...]* -niŋ|NEG (444)
[...]* lut|tell [...]* na|PTCL (403)
[...]* mo|PTCL lut|tell [...]* na|PTCL (381)
[...]* -ko|GEN (335)
[...]* -u|3P -kV|NPST (320)
[...]* mo|PTCL [...]* -?|EMPH na|PTCL (312)
[...]* kha|FOC (306)
[...]* -ŋs|PRF -e|PST (293)
[...]* pho|REP (290)
[...]* -e|PST (2221)
[...]* -no|NPST (1617)
[...]* -a|IMP [...]* na|PTCL (911)
[...]* -a|PST [...]* -e|PST (805)
[...]* -a|IMP na|PTCL (674)
[...]* -?|EMPH na|PTCL (619)
[...]* -a|IMP -?|EMPH na|PTCL (612)
[...]* -a|IMP [...]* -a|IMP (552)
[...]* lo|PTCL (486)
[...]* o|EMPH (461)
[...]* -ŋ|1sA (444)
[...]* -u|3P [...]* na|PTCL (394)
[...]* naŋ|PTCL (375)
[...]* ta|FOC (333)
[...]* mo|PTCL [...] -a|IMP [...] na|PTCL (317)
[...]* -ce|ns (306)
[...]* -ni|DIR (300)
[...]* a-|2 [...]* -e|PST (292)
[...]* -hatt|TEL -e|PST (289)

[...]* e|OR (276)
[...]* -ce|d (270)
[...]* mo|PTCL [...] -a|IMP -ʔ|EMPH na|PTCL (268)
[...]* -u|3P [...]* -e|PST (265)
[...]* ba|DEM.PROX (255)
[...]* -ŋa|ERG (234)
[...]* -kV|NPST -ŋ|1sA (223)
[...]* -u|3P -kV|NPST -ŋ|1sA (215)
[...]* yuŋ|be -no|NPST (205)
[...]* -u|3P -ŋ|1sA (194)
[...]* -a|PST -ŋs|PRF -e|PST (192)
[...]* -ŋa|ERG [...]* -e|PST (185)
[...]* them|what (175)
[...]* -hē|ePST (171)
[...]* -ŋs|PERF -e|PST (161)
[...]* khat|go -e|PST (153)
[...]* -no|NPST -nɨŋ|NEG (149)
[...]* -u|3P (138)
[...]* thitta|one (135)
[...]* -ŋa|1sS/P -ʔā|1sNPST (131)
[...]* huĭ|DEM (126)
[...]* -hatt|TEL [...]* -e|PST (122)
[...]* -no|NPST [...]* na|PTCL (117)
[...]* -e|PST aŋ|PTCL (110)
[...]* -u|3P [...]* -ŋ|1sA (272)
[...]* a-|2 [...] -no|NPST (268)
[...]* -a|PST -hatt|TEL -e|PST (267)
[...]* -kho|IMP (264)
[...]* lo|ok (252)
[...]* -u|3P -kho|IMP (230)
[...]* -ma|INF (215)
[...]* ho|PTCL (207)
[...]* -pe?|LOC (197)
[...]* manchi|not (192)
[...]* yaŋ|ADD (187)
[...]* a-|2 [...]* -kV|NPST (178)
[...]* -ʔā|1sNPST (172)
[...]* hola|probably (165)
[...]* huŋgo|DEM (153)
[...]* -a|IMP [...]* -ʔ|EMPH (151)
[...]* mo|PTCL (145)
[...]* kina|SEQ (135)
[...]* -ma|INF [...]* -no|NPST (135)
[...]* -ʔ|EMPH [...]* -ʔ|EMPH (130)
[...]* -e|PST [...]* na|PTCL (125)
[...]* abo|now (120)
[...]* khat|go -a|IMP (113)
[...]* -ko|GEN [...]* na|PTCL (110)

Table A.3: Utterance-medial patterns

[...]* -a|IMP [...]* (5000)
[...]* -e|PST [...]* (2637)
[...]* -no|NPST [...]* (1951)
[...]* -ʔ|EMPH [...]* (1681)
[...]* -ko|GEN [...]* (1459)
[...]* -kV|NPST [...]* (1288)
[...]* mo|PTCL [...]* (1114)
[...]* -ma|INF [...]* (1083)
[...]* -u|3P [...]* (4276)
[...]* -a|PST [...]* (2030)
[...]* a-|2 [...]* (1907)
[...]* khat|go [...]* (1521)
[...]* -ce|d [...]* (1414)
[...]* -a|IMP -ʔ|EMPH [...]* (1125)
[...]* -hatt|TEL [...]* (1097)
[...]* na|PTCL [...]* (1032)

[...]* ta|FOC [...]* (1006)

[...]* -ce|ns [...]* (976)

[...]* -a|PST [...]* -e|PST [...]* (896)

[...]* lo|PTCL [...]* (885)

[...]* -a|IMP [...]* -a|IMP [...]* (838)

[...]* yaŋ|ADD [...]* (797)

[...]* numdi|do [...]* (713)

[...]* ci|eat [...]* (701)

[...]* u-|3sPOSS [...]* (677)

[...]* -u|3P [...]* -u|3P [...]* (639)

[...]* pit|give [...]* (583)

[...]* -ŋa|1sS/P [...]* (570)

[...]* na-|3>2 [...]* (547)

[...]* -niŋ|NEG [...]* (517)

[...]* -a|IMP -ce|d [...]* (496)

[...]* -ŋ|1sA [...]* (492)

[...]* -a|PST -ŋs|PRF [...]* (490)

[...]* -kho|IMP [...]* (478)

[...]* lis|be [...]* (471)

[...]* them|what [...]* (443)

[...]* -ŋs|PERF [...]* (404)

[...]* ta|PTCL [...]* (401)

[...]* u-|3nsS/A [...]* (382)

[...]* -ŋs|PRF -e|PST [...]* (372)

[...]* -ce|d -a|IMP [...]* (364)

[...]* lut|tell -a|IMP [...]* (362)

[...]* kina|SEQ [...]* (338)

[...]* yuŋ|sit [...]* (333)

[...]* -dhend|TEL -u|3P [...]* (327)

[...]* -ni|DIR [...]* -a|IMP [...]* (324)

[...]* mo|PTCL [...]* -?|EMPH [...] (318)

[...]* -na|1>2 [...]* (313)

[...]* ni|PTCL [...]* (311)

[...]* -ko|NMLZ [...]* (306)

[...]* khatt|take [...]* (302)

[...]* -u|3P -kV|NPST [...]* (1001)

[...]* -ŋa|ERG [...]* (918)

[...]* -ni|DIR [...]* (886)

[...]* -ŋs|PRF [...]* (851)

[...]* ba|DEM.PROX [...]* (808)

[...]* -a|PST -hatt|TEL [...]* (722)

[...]* kha|PTCL [...]* (702)

[...]* -dhend|TEL [...]* (698)

[...]* -pe?|LOC [...]* (642)

[...]* mo|PTCL [...] -a|IMP [...]* (585)

[...]* akka|1s [...]* (576)

[...]* -i|LOC [...]* (565)

[...]* lut|tell [...]* (545)

[...]* a-|2 [...]* -u|3P [...]* (498)

[...]* yuŋ|be [...]* (493)

[...]* mai-|NEG [...]* (491)

[...]* mett|do [...]* (488)

[...]* i-|2sPOSS [...]* (475)

[...]* mo|PTCL lut|tell [...]* (457)

[...]* -u|3P -kho|IMP [...]* (422)

[...]* -a|N.NTVZ [...]* (402)

[...]* yo|DEM.ACROSS [...]* (393)

[...]* -na|NA [...]* (381)

[...]* hana|2s [...]* (366)

[...]* bai?|DEM.PROX [...]* (364)

[...]* -a|PST [...]* -a|PST [...]* (356)

[...]* khoŋs|play [...]* (337)

[...]* a-|1sPOSS [...]* (329)

[...]* kha|FOC [...]* (327)

[...]* -u|3P -ŋs|PRF [...]* (318)

[...]* -ŋa|ERG [...]* -u|3P [...]* (317)

[...]* kat|come.up [...]* (312)

[...]* huŋgo|DEM [...]* (306)

[...]* -ce|3nsP [...]* (305)

[...]* -e|PST -?|EMPH [...]* (300)

[...]* ma|mother [...]* (294)

[...]* -u|3P [...]* -e|PST [...]* (292)

[...]* khat|go -a|IMP [...]* (283)

[...]* a-|2 [...]* -e|PST [...]* (279)

[...]* mo|PTCL [...] -a|IMP -ʔ|EMPH [...] (272)

[...]* hokke|where [...]* (254)

[...]* -t|NEG [...]* (250)

[...]* pa|father [...]* (245)

[...]* -a|PST -dhend|TEL [...]* (244)

[...]* -a|PST -ŋs|PRF -e|PST [...]* (238)

[...]* -i|1/2pS/P [...]* (231)

[...]* a-|2 khat|go [...]* (227)

[...]* -ni|DIR khat|go [...]* (221)

[...]* cekt|say [...]* (216)

[...]* -u|3P [...]* -ŋ|1sA [...]* (214)

[...]* -tha|NEG.IMP [...]* (210)

[...]* a-|2 [...] -a|PST [...]* (204)

[...]* -i|p [...]* (196)

[...]* -si|PURP [...]* (192)

[...]* -e|V.NTVZ [...]* (189)

[...]* -a|PST [...]* -ŋs|PRF [...]* (188)

[...]* e|OR [...]* (185)

[...]* thams|fall.down [...]* (184)

[...]* khat|go -a|PST [...]* (183)

[...]* yuŋ|be -no|NPST [...]* (182)

[...]* catt|hit [...]* (180)

[...]* -a|IMP [...]* -ʔ|EMPH [...]* (178)

[...]* pho|REP [...]* (175)

[...]* -hatt|TEL -a|PST [...]* (171)

[...]* ti|come [...]* (167)

[...]* hit|be.able [...]* (166)

[...]* sa|who [...]* (162)

[...]* khatt|take -u|3P [...]* (160)

[...]* -kV|NPST -ŋ|1sA [...]* (158)

[...]* tis|put [...]* (156)

[...]* -a|IMP -ce|d -a|IMP [...]* (294)

[...]* -u|3P -ŋ|1sA [...]* (285)

[...]* -hatt|TEL -e|PST [...]* (281)

[...]* thap|come.level [...]* (279)

[...]* -a|PST -hatt|TEL -e|PST [...]* (264)

[...]* u-|3A [...]* (253)

[...]* mo|DEM.down [...]* (246)

[...]* a-|2 [...] -no|NPST [...]* (244)

[...]* -a|PST -ŋs|PERF [...]* (240)

[...]* cekt|speak [...]* (231)

[...]* thapt|bring.across [...]* (230)

[...]* -u|3P [...] -a|IMP [...]* (224)

[...]* a-|2 [...]* -kV|NPST [...]* (219)

[...]* -sa|OBL [...]* (215)

[...]* huĭ|DEM [...]* (214)

[...]* -a|IMP [...]* na|PTCL [...]* (205)

[...]* -a|IMP na|PTCL [...]* (199)

[...]* -ko|GEN u-|3sPOSS [...]* (195)

[...]* lut|say [...]* (190)

[...]* khutt|bring.sth.for.sb [...]* (188)

[...]* -ko|GEN [...]* -u|3P [...]* (187)

[...]* thitta|one [...]* (184)

[...]* -u|3P -dhend|TEL [...]* (184)

[...]* -le|RESTR [...]* (183)

[...]* -u|3P -ce|3nsP [...]* (180)

[...]* copt|see [...]* (178)

[...]* -no|NPST -nɨŋ|NEG [...]* (176)

[...]* kuŋs|come.down [...]* (173)

[...]* -ce|ns [...]* -u|3P [...]* (169)

[...]* -ma|INF [...]* -u|3P [...]* (166)

[...]* numd|do -a|IMP [...]* (163)

[...]* aŋ|what [...]* (161)

[...]* yuŋ|sit -a|IMP [...]* (158)

[...]* -a|PST -hatt|TEL -a|PST [...]* (158)

[...]* yuŋs|keep [...]* (155)

[...]* khat|go -e|PST [...]* (154)

[...]* -niŋ|COM [...]* (153)

[...]* abo|now [...]* (149)

[...]* -a|IMP -dhend|TEL [...]* (144)

[...]* -ce|d -kV|NPST [...]* (142)

[...]* -hatt|TEL -u|3P [...]* (141)

[...]* -u|3P -ŋs|PERF [...]* (139)

[...]* lond|come.out [...]* (138)

[...]* mett|do -u|3P [...]* (135)

[...]* nis|know [...]* (131)

[...]* -?|EMPH na|PTCL [...]* (127)

[...]* -ŋa|ERG na-|3>2 [...]* (126)

[...]* -a|PST [...] -u|3P [...]* (121)

[...]* -ŋa|1sS/P -?ă|1sNPST [...]* (116)

[...]* -u|3P -kV|NPST -ŋ|1sA [...]* (154)

[...]* mo|DEM.DOWN [...]* (152)

[...]* -u|3P -dhend|TEL -u|3P [...]* (145)

[...]* ta|FOC [...]* -u|3P [...]* (143)

[...]* -?ă|1sNPST [...]* (142)

[...]* -a|IMP [...]* -u|3P [...]* (141)

[...]* manchi|not [...]* (139)

[...]* a-|2 [...] -u|3P -kV|NPST [...]* (135)

[...]* -hatt|TEL [...] -ŋs|PRF [...]* (132)

[...]* -a|PST [...] -a|PST -ŋs|PRF [...]* (131)

[...]* nunu|baby [...]* (127)

[...]* hap|cry [...]* (124)

[...]* -a|PST -hatt|TEL [...] -ŋs|PRF [...]* (117)

[...]* -hatt|TEL -a|PST -ŋs|PRF [...]* (107)

Appendix B

Summary in German

Diese Arbeit beschäftigt sich mit dem methodologischen Problem der Mustererkennung in Sprachkorpora. Das wissenschaftliche Interesse an diesem Problem ist vor allem in der Spracherwerbsforschung begründet. Die Arbeit diskutiert den aktuellen Forschungsstand zum Sprachwerwerb und erklärt inwiefern bestehende Methoden zur Mustererkennung unzureichend sind. Daraufhin wird ein auf mathematischem Modell begründetes Verfahren zu Musteranalyse vorgestellt. Anschließend wird dieses Verfahren eigesetzt um ein Spracherwerbskorpus der Kirantisprache Chintang auszuwerten. Die Ergebnise werden zusammengefasst und mit anderen Arbeiten im gleichen Gebiet verglichen. Es wird das Fazit gezogen, dass unterschiedliche nicht-verwandte Sprachen viele Ähnlichkeiten in Bezug auf die statistischen Muster im Sprachgebrauch aufzeigen.

Bibliography

A. Perfors, J. Tenenbaum & T. Regier (2006) Poverty of the Stimulus? A Rational Approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vancouver, Canada.

Abbot-Smith, K. & M. Tomasello (2006) Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* **23**.

Bickel, B., G. Banjade, M. Gaenszle, E. Lieven, N. P. Paudyal, I. P. Rai, M. Rai, N. K. Rai, & S Stoll (2005a) Triplication and ideophones in Chintang. In *Contemporary issues in Nepalese linguistics*, Y. P. Yadava, ed., Linguistic Society of Nepal, Kathmandu.

Bickel, B., G. Banjade, M. Gaenszle, E. Lieven, N. P. Paudyal, I. P. Rai, M. Rai, N. K. Rai, & S Stoll (2005b) Worshipping the King God. A preliminar y analysis of Chintang ritual language in the invocation of Rajdeu. In *Contemporary issues in Nepalese linguistics*, Y. P. Yadava, ed., Linguistic Society of Nepal, Kathmandu.

Bickel, B., G. Banjade, M. Gaenszle, E. Lieven, N. P. Paudyal, I. P. Rai, M. Rai, N. K. Rai, & S. Stoll (2007) Free prefix ordering in Chintang. *Language* **83**.

Bohannon, John N. & Laura B Stanowicz (1988) The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology* **24**(5).

Bowerman, M. (1988) How Do Children Avoid Constructing an Overly General Grammar in the Absence of Feedback about What Is Not a Sentence? In *Explaining language universals*, J. A. Hawkins, ed., Oxford: Basil Blackwell.

Cameron-Faulkner, T., E. Lieven, & M. Tomasello (2003) A construction based analysis of child directed speech. *Cognitive Science* **27**: 843–873.

Cartwright, T.A. & M.R. Brent (1997) Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition* **63**.

Chomsky, N. (1956) Three models for the description of language. *Information Theory, IEEE Transactions on* **2**(3): 113–124.

Chomsky, N. (1980a) On cognitive structures and ther development: a reply to Piaget. In *Language and learning: the debate between Jean Piaget and Noam Chomsky*, Massimo Piattelli-Palmarini, ed., Routledge.

Chomsky, N. (1980b) *Rules and Representations*. Columbia University Press.

Chomsky, Noam (1995) *The Minimalist Program*. Cambridge: MIT Press.

Crain, S. (1991) Language acquisition in the absence of experience. *Behavioral and Brain Sciences* **14**.

Elman, J. L. (2002) Generalization from sparse input. In *Proceedings of the 38th Annual Meeting of the Chicago Linguistic Society. Vol. 2: The panels*, Andronis M., E. Debenport, A. Pycha, & Yoshimura K., eds., Chicago: CLS.

Elman, Jeffrey L. (1993) Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition* **48**(1): 71–9.

Gerken, Louann, Rachel Wilson, & William Lewis (2005) Infants can use distributional cues to form syntactic categories. *Journal of Child Language* **32**(02): 249–268.

Gold, E. Mark (1967) Language identification in the limit. *Information and Control* **10**(5): 447–474.

Gomez, R.L. & Gerken L. (1999) Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* **70**.

Gordon, Peter (1990) Learnability and feedback. *Developmental Psychology* **26**(2).

Grimshaw, J. (1981) Form, function, and the language acquisition device. In *The logical problem of language acquisition*, C.L. Baker & J. McCarthy, eds., MIT Press.

Harris, Z.S (1964) Distributional structure. In *The Structure of Language*, J.A. Fodor & J.J. Katz, eds., Prentice Hall.

Hope, A. C. A. (1968) A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society: Series B* **30**: 582–598.

Hornstein, N. & D. Lightfoot (1981) Introduction. In *Explanation in linguistics: the logical problem of language acquisition*, Longman, London ; New York.

Jusczyk, Peter W. (1997) *The discovery of spoken language*. Cambridge, Mass.: MIT Press.

Kelly, M. H. (1992) Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review* **99**(2).

Kelly, M. H. (1996) The role of phonology in grammatical category assignments. In *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, J. L. Morgan & K. Demuth, eds., Mahwah, New Jersey: Erlbaum.

Kiss, G.R. (1973) Grammatical word classes: a learning process and its simulation. *Psychology of Learning and Motovation* **7**.

Kracht, Marcus (2003) The Mathematics of Language. In *Studies in Generative Grammar*, 3–11017620.

Levene, Howard (1960) Robust tests for equality of variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Ingram Olkin & Harold Hotelling, eds., Stanford University Press.

Lieven, E. V. M., J. M. Pine, & G. Baldwin (1997) Lexically-based learning and early grammatical development. *Journal of Child Language* **24**(01): 187–219.

Maratsos, M. P. & M. A. Chalkley (1980) The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In *Children's language*, vol. 2, K. E. Nelson, ed., Gardner.

Mintz, T.H. (2002) Category induction from distributional cues in an artificial language. *Memory & Cognition* **30**(5).

Mintz, T.H. (2003) Frequent frames as a cue for grammatical categories in child directe speech. *Cognition* **91**: 91–117.

Monaghan, Padraic, Morten H. Christiansen, & Nick Chater (2007) The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology* **55**(4): 259 – 305.

Pinker, S. (1984) *Language learnability and language development*. Harvard university press.

Pinker, S. (1987) The bootstrapping problem in language acquisition. In *Mechanisms of language acquisition*, B. MacWhinney, ed., Hillsdale, NJ: Lawrence Erlbaum Associates.

Pullum, G. K. & B. C. Scholz (2002) Empirical assessment of stimulus poverty arguments. *The Linguistic Review* **19**(1-2): 9–50.

Redington, M., N. Chater, & S. Finch (1998) Distributional Information: A powerful Cue for Aquiring Syntactic Categories. *Cognitive Science* **22**(4): 425–469.

Rohde, Douglas & David C. Plaut (1999) Language Acquisition in the Absence of Explicit Negative Evidence: How Important is Starting Small? *Cognition* **72**: 67–109.

Royston, Patrick (1995) Remark AS R94: A remark on Algorithm AS 181: The W test for normality. *Applied Statistics* **44**.

S., Crain & Nakayama M. (1987) Structure Dependence in Grammar Formation. *Language* **63**(3).

Saffran, J.R, R.N Aslin, & E.L. Newport (1996) Statistical Learning by 8-Month-Old Infants. *Science* **13**.

Saussure, Ferdinand de, Charles Bally, Albert. Riedlinger, & Albert. Sechehaye (1960) *Course in general linguistics / Ferdinand de Saussure ; edited by Charles Bally and Albert Sechehaye, in collaboration with Albert Reidlinger ; translated from the French by Wade Baskin*. Owen, London [England] :, 1st british commonwealth ed. ed.

Scholz, B. C. & G. K Pullum (2002) Searching for arguments to support linguistic nativism. *The Linguistic Review* **19**(1-2): 9–50.

Shi, R., J.L. Morgan, & P. Allopenna (1998) Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language* **25**.

Shieber, Stuart M. (1985) Evidence against the context-freeness of natural language. *Linguistics and Philosophy* **8**(3): 333–343.

Shinohara, T. (1994) Rich classes inferable from positive data: length-bounded elementary formal systems. *Information and computation* **108**.

Sokolov, J. L. & C. E. Snow (1994) The changing role of negative evidence in theories of language development. In *Input and interaction in language acquisition*, C. Gallaway & B. J. Richards, eds., Cambridge: Cambridge University Press.

Stoll, S., K. Abbot-Smith, & E. Lieven (2009) Lexically Restricted Utterances in Russian, German and English Child-Directed Speech. *Cognitive Science* **33**: 75–103.

Stoll, S., B. Bickel, E. Lieven, G. Banjade, T. N. Bhatta, M. Gaenszle, N. P. Paudyal, J. Pettigrew, I.P. Rai, M. Rai, & N.K. Rai (2008) Nouns and verbs in Chintang: children's usage and surrounding adult speech.

Tomasello, M. (2003) *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Selbstständigkeitserklärung

Ich bin damit einverstanden, dass meine Magisterarbeit in der Bibliothek öffentlich eingesehen werden kann. Die Urheberrechte müssen gewahrt werden. Die Arbeit enthält keine personenbezogenen Daten.

Datum Unterschrift

Hiermit versichere ich, dass ich die Arbeit in allen Teilen selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe.

Datum Unterschrift