1 Introduction

1.1 Languages of South Asia

South Asian languages (hereafter, SALs), comprising four major language families (Austro-Asiatic with two sub-branches – Mon-Khmer and Munda – Dravidian, Indo-Aryan and Tibeto-Burman), are rich in linguistic diversity, providing a great opportunity for investigation.[1] Though estimates differ as to how many languages are spoken in the subcontinent, it is generally agreed that the number runs to more than 100.

This book focuses on the syntactic typology of SALs in general, and the high degree of syntactic convergence in particular, with special reference to the notion of "India as a linguistic area" (Emeneau 1956; Masica 1976; see chapter 2 for details).

The study of syntactic typology has gained enormous importance in the last twenty-five years with emphasis on data and analysis of unrelated and genetically distinct languages. It enables us to understand the intricate nature of the Universal Grammar (UG) and parametric variation. Universal Grammar is a set of principles that are commonly shared by all languages of the world. According to Chomsky (1975: 118), a general principle found in a language "belongs to universal grammar, as part of 'pre-existent' knowledge that makes learning possible." Languages do not differ from each other in *innumerable* or *random ways*; rather, they differ from each other in terms of a limited set of parameters relating to the subsystems of a formal grammar. Based on data from the four major South Asian language families mentioned, we wish to examine the principles that SALs share with other languages in terms of the UG, and the parameters of syntax and morphology according to which they differ. We wish to demonstrate how a formal theory of language such as the Government and Binding framework allows us to capture the commonalities and differences amongst languages in a principled manner.

Our work also focuses on South Asia as a "linguistic area." Languages in the subcontinent share a number of features at the phonological, syntactic and semantic levels. However, they do differ from each other in their own unique

ways, demonstrating that each language retains its identity in spite of intense contact with other languages either due to multilingualism, or areal contact, or both. The similarities found in SALs are due to the fact that all the languages, except for a few such as Khasi, Rymbai and Pnar (Mon-Khmer), are verb-final. As one would expect, verb-final languages have several features in common. Our work further demonstrates how languages that are contiguously located in an area can affect each other in a radical and in a principled manner.

1.2 Aim of the book

The aim of this book is to provide a typology of major constructions selected from the principal components of grammar (such as relative clauses, case, agreement, anaphora, complementation, conjunctive participles, etc.) with a view to arriving at an understanding of the intricate nature of such constructions and to see how such analysis can enrich our understanding of Universal Grammar and parametric variation. This will be accomplished first by presenting, in a theory-neutral fashion, the basic facts of each of the topics chosen, and then by providing an explanation of the phenomena in the modular approach of the Government and Binding framework. Phrase structure, the role of Case, binding, thematic relations, control theory and movement as the various modules of the Government and Binding framework, and significant notions such as government and c-command, have come into focus in the last three decades. Never before has the study of language universals and parametric variation received as much attention as it has since the eighties. The topics chosen and the issues discussed are intended to enable the reader to get a perspective on the nature of language variation, and the universal principles involved, using the concepts of a formal theory of grammar, and also to define exactly the features involved in convergence.

While providing an overall typology of the aspects of SALs, care was taken in each chapter to focus on the issues that concern individual languages and the language family to which they belong. For example, there has been a considerable debate over the issue of *one finite verb per sentence* and the Strict OV Constraint in Dravidian. This issue relates to the occurrence of the embedded complement clauses with the marker -*ō*, and the relative-correlative constructions with the same marker -*ō*. The nature of the finiteness phenomenon has implications for the nature of contact and convergence between Dravidian and Indo-Aryan (see Hock 2005 for the most recent discussion). In the discussion of conjunctive participles, complementation and relative clauses, this issue has been focused upon and discussed.

1.3 Linguistic theory, language universals and language typology

Language universals and parametric variation have been the focus of study for the last thirty years, and significant insights have been gained in our understanding of the nature of human language. An explanatorily adequate theory needs to be supported and substantiated by empirical facts, while empirical facts in isolation do not have much relevance unless they are properly explained in an appropriate theoretical framework. The last three decades have witnessed an enormous emphasis on presenting evidence based on empirical data so that the foundations prepared for theoretical claims are well laid (Bhaskararao and Subbarao 2004: ix) – for instance, as in the case of lexical anaphors (Lust, Wali, Gair and Subbarao 2000) or non-nominative subjects (Bhaskararao and Subbarao 2004). Thus, to build an explanatorily adequate theory, it is essential to have a sound database. A fruitful combination of theory and data alone will yield the desired results. Language typology attempts to study linguistic patterns and language variation, and arrive at generalizations from cross-linguistic comparison (Croft 1990). One of the goals of typological studies should be to advance our understanding of central theoretical questions of language. Hence, typology must focus on various phenomena of current theoretical and empirical relevance, and must attempt to provide solutions supported by rich empirical data and valid arguments. We feel that a *symbiotic relationship between linguistic theory and language typology* which primarily emphasizes choosing data from a wide variety of languages and analyzing them in a theoretical framework is thus a desideratum (Subbarao 1999). A similar view is articulated in Cinque (2007: 93), who does not consider "('formal') linguistic theory and linguistic typology as two separate approaches." Our work in this volume attempts to combine, as Baker and McCloskey (2007: 291) succinctly put it, "*a motivated degree of abstractness in the analysis* of particular languages, as is typical in formal linguistics, with an interest in *sampling widely from a range of languages*, as is typical in typological studies [emphasis added]." Baker (2001) labels such an approach as the "the middle way."

In a sense, this book is a manifestation of what Baker (2008) calls Formal Generative Typology, aiming to find out "what properties of natural human languages are genuinely universal," "what properties of natural human languages vary from one human language to another" and "which aspects of variation are patterned, systematic, and grammatical in nature, and which aspects of variation are random, idiosyncratic, and lexical in nature."

While it is the linguistic theory that enables linguists to formulate a hypothesis in a given framework, helps them to look for the relevant data, offers them a direction in which to look, it is the data that not only substantiate the claims made in the theory, but also provide a window to look for inadequacies,

counterexamples to the theoretical claims made, enabling linguists to revise and enrich the theoretical base, if necessary. Hence, an active interaction between data and theory is a must. Any theory that is proposed without a sound theoretical base and any theoretical claims made regarding the nature of human language without a strong database will not be explanatorily adequate.

We attempt to demonstrate with our findings that it is the linguistic theory with a sound theoretical base that works as a powerful tool in explicating commonness and variation in languages. This book is based on a rich set of data from a larger number of languages and these data have been analyzed uniformly through the perspective of one theory.

1.4 Inductive and deductive approaches to language analysis

Most of the studies in linguistic typology have been done either purely in rationalist (deductive) or in empiricist (inductive) approaches. It was in some studies in relational grammar, a study by Keenan and Comrie (1977) and some studies in the Government and Binding framework that attempts were made to strike a reasonable balance between theory and data in analysis.

Though not overtly spelt out, there have been two distinct approaches to the study of language universals: inductive and deductive. While inductivists believe in drawing their generalizations based on data collected from a wide range of languages, deductivists may arrive at these constraints based on a metalinguistic theory, drawing their evidence from a single language or a limited number of languages. Scholars who are opposed to a totally deductive approach argue that "it is necessary to have data from a wide range of languages" (Comrie 1981: 4) to carry out research on language universals. Thus, there is a great emphasis on arriving at language universals on the basis of "concrete rather than abstract analyses."

Deductivists, in contrast, argue that the theory of grammar has the capacity to predict language universals. It was generally agreed that the study of even a single language is sufficient to predict language universals. Chomsky (1975: 118) points out: "the principles that appear to have explanatory adequacy for English are the principles of universal grammar." Thus, "[a] great deal can be learned about UG from the study of a single language" (Chomsky 1981: 6). Coopmans (1983: 458), too, articulates a similar view: "The universals of generative grammar have a different basis. If we assume that these principles are universally available, in the sense that they are genetically encoded, then by a detailed study of one particular language, we will be able to discover *some* properties of UG and thus, deepen our insights on the acquisition process [emphasis added]."

The deductivist favors innateness as the explanation for language universals. In view of the methodology followed and the choice of data collected for

linguistic inquiry, we argued elsewhere that the deductivist is a microinductivist and the inductivist is a microdeductivist and, therefore, a strict dichotomy between the inductive (empiricist) and deductive (rationalist) approaches cannot be maintained (Subbarao and Saxena 1987; Subbarao 1999). We proposed an integrated approach that combines these two. This can be summed up as:

Limited Induction Base → Deduction → Induction

(See Subbarao and Saxena 1987 for details.)

Thus, to build an explanatorily adequate theory, it is essential to have a sound database. A fruitful conflation of theory and data alone will yield the desired results. According to Lust *et al*. (2000: 1): "Without basic theory, the fundamental questions of the nature of human competence for a language cannot be addressed. However, in the absence of real language data, proposed answers may not be relevant to the real questions." Lust *et al.* (2000) label their approach as "principled typology." Consequently, it should be the concern of the theoretical linguist as well as the typologist to combine theory and empirical language data.

With this goal in mind, each chapter in this book strikes a balance between theoretical assumptions and the issues that arise from the data presented from sixty languages belonging to the four language families of the South Asian subcontinent.

In each chapter, we provide a detailed analysis of the data in theory-neutral terms first. It is envisaged that any researcher, irrespective of their theoretical background, can understand such analysis. Once this is done, points in the theoretical framework of the Principles and Parameters approach are analyzed (Chomsky and Lasnik 1995).

1.5 Relevance of linguistic theory: some illustrations

We attempt to demonstrate that it is the linguistic theory with a sound theoretical base that works as a powerful tool to explicate differences in languages. We discuss below some specific cases from SALs where theory provides an explanation to account for the facts.

(i) The first example comes from the occurrence of nominal anaphors and the verbal anaphor. SALs have two forms of the nominal anaphor – simplex and complex – and two forms of the verbal anaphor – verbal reflexive and verbal reciprocal. The verbal anaphor is universally monomorphemic (a simplex form). The complex nominal anaphor is either a partially or fully reduplicated form of the simplex form.

Except in Marathi (Indo-Aryan [IA]), long-distance binding is blocked when the complex form of the *nominal anaphor* occurs in the embedded clause. We

demonstrate that this fact can be explained by invoking *Head to Head movement* (see chapter 3 for details).

When the *verbal anaphor* occurs in the embedded clause, long-distance binding is blocked in all the languages.

Thus, the occurrence of either a nominal complex anaphor or a verbal anaphor or both blocks long-distance binding.

Leaving the case of Marathi (IA) aside for a moment, this can be neatly formulated in terms of the following parameter.

> *Parameter 1:*
> [−/+ complex nominal anaphor] / [−/+ verbal anaphor] directly correlates with [+/− long-distance binding].

That is, the occurrence of the complex nominal anaphor and verbal anaphor are *negatively specified* for the phenomenon of long-distance binding.

(ii) The second example concerns the occurrence of the verbal reflexive and the locative PP. A locative PP may either be a subcategorized or a non-subcategorized argument of a predicate. When an anaphor occurs in a subcategorized locative PP position, the occurrence of the verbal reflexive is obligatory – otherwise, it is not (Lust *et al*. 2000). Thus, the notion of *subcategorization*[2] helps us in providing an explanation of the occurrence of the verbal reflexive.

This can be formulated in terms of the following parameter.

> *Parameter 2:*
> [+/− the occurrence of a verbal anaphor] directly correlates with [+/− the occurrence of a nominal anaphor in a subcategorized locative PP position] (see chapter 3 for details).

That is, the occurrence of a nominal anaphor in a subcategorized locative PP position is *positively specified* for the occurrence of a verbal anaphor.

(iii) The third example concerns the case of the occurrence of initial complementizer (IC) in complement clauses in verb-final languages and wide-scope reading of wh-expressions.

Word order typology and the Head Direction parameter help us predict wide- and narrow-scope interpretations of wh-expressions in complement clauses. Bayer (2001) argues that in verb-final Indo-Aryan languages, wide-scope reading of wh-expressions in IC clauses is *blocked*. This is due to the fact that the IC clause falls outside the pattern of *argument licensing* in head-final languages. Our study substantiates the claim made in Bayer (2001).

In contrast, verb-final languages with a final complementizer (FC) which is a quotative permit wide-scope reading of the wh-expressions. The various parameters that affect the scope of a question word in complement clauses are discussed and ten parameters are suggested (see chapter 6 for details).

This can be formulated in terms of the following parameter.[3]

> *Parameter 3:*
> [+/− initial complementizer] in verb-final languages directly correlates with [−/+ wide-scope reading of the wh-expression].

That is, wide-scope reading of the wh-expression is *negatively specified* for the feature of initial complementizer.

(iv) The fourth example comes from relative clauses in SALs. Some Tibeto-Burman languages (Angami [a.k.a. Tenyidie], Sema, Mizo, Manipuri, Sangtam and Konyak) have both externally headed and internally headed relative clauses. In externally headed relatives, using a gap rather than movement in many SALs, a comitative PP cannot head a gap (infinitival) relative clause. In those languages in which it can, we have demonstrated that a *thematic relation* needs to be established between the predicate of the embedded clause and the comitative head of the externally headed and internally headed relative clause, either in terms of an overt case marker or some verbal marker/clitic in the embedded verb to indicate accompaniment. A similar phenomenon is found in Khasi and Pnar (Mon-Khmer) too. Though the notion of thematic relations has run into rough weather in recent years (see Newmeyer 2007: 140–141) and it is difficult to agree on which "thematic roles exist" and how "to independently justify the assignment of noun phrases to thematic roles in particular sentences" (Dowty 1989: 70, as quoted in Newmeyer 2007: 140–141) in an objective manner, our analysis demonstrates that the phenomenon of incorporation of either an adverb such as *together* or a postposition or a reciprocal marker as a group marker in the embedded verb enables the comitative PP to meet the Thematic Eligibility Condition (TEC) proposed in this study, and a comitative PP can consequently head a non-finite relative clause. Thus, the TEC needs to be included in the grammar in order to explicate the occurrence or non-occurrence of a comitative PP as head in the gap (infinitival) relative (see section 8.7 for a detailed discussion).[4]

This can be formulated in terms of the following parameter.

> *Parameter 4:*
> [+/− the occurrence of a comitative adposition with the head NP / an adverb denoting *together* along with a verbal clitic / an incorporated comitative adposition in the embedded verb] directly correlates with [+/− the occurrence of a comitative PP as head] in the gap relative.

That is, the occurrence of a comitative PP is *underspecified* for the occurrence of a comitative PP as head.

(v) The fifth example concerns the accusative case-marking in nominative–accusative sentences and the non-nominative (dative/genitive subject) constructions. The "objects" are differentially marked depending upon features such

as *specificity* and *animacy*. Thus, with regard to differential object marking (DOM) in SALs, when the noun phrase denoting specificity/animacy occurs, the accusative marker may occur even though the predicate is [−transitive] in the dative/genitive subject construction (DSC) in Bangla (IA), Malayalam and Tamil (DR). When the predicate is [+transitive] in the nominative subject construction, the object is accusative case-marked when it denotes specificity/animacy. Based on these facts, we propose the following parameter to account for this variation.

> The Differential Object Marking (DOM) parameter: when the noun phrase is accusative case-marked, the object marker is either associated with animacy/specificity depending on transitivity in the [−NNS construction],[5] or purely animacy/specificity independent of transitivity in the [+NNS construction], as in parameter 5.
>
> *Parameter 5:*
> The occurrence of the accusative case marker denoting animacy/specificity directly correlates with [−/+transitivity] in the [+/−NNS construction] (see chapter 5 for a detailed discussion).

We present further arguments in this work to demonstrate that typology without theory, and theory without typology, are incomplete, and this, we hope, will further strengthen our claim that there is a *symbiotic relation* between the two.

1.6 The framework

In each chapter, after the presentation of data, there is an analysis of the significant points in the theoretical framework of the Principles and Parameters approach.

1.6.1 The Principles and Parameters approach

Languages appear to differ from each other in an innumerable number of ways. According to the Principles and Parameters approach (Chomsky and Lasnik 1995), this variation is not random, and is due to a specific set of parameters, which are *limited* in number. It is the interaction of the parameters with the universal set of principles that are common to all languages that constitutes human language.

1.7 The data

Data for the volume are drawn from the following SALs. All the languages are verb-final in sentence structure, except for Khasi (Mon-Khmer), in which

the verb occurs in the medial position, and Kashmiri (Indo-Aryan), in which the verb occurs in the V_2 position, though it is underlyingly verb-final (see chapter 2 for details).

A list of the languages from the four South Asian language families from which our work draws its data is given below:

ia *Indo-Aryan*
- Assamese
- Bangla
- Gujarati
- Hindi-Urdu
- Kashmiri
- Konkani
- Maithili
- Marathi
- Nepali
- Oriya
- Punjabi
- Eastern Shina
- Shina of Gultari
- Shina of Skardu
- Sinhala
- Swat-Dir Kohistani
- Torwali

ib *Contact languages*
- Bhalavali Bhasha (a.k.a. Bhalawali Marathi)
- Dakkhini (Hindi-Urdu)
- Eastern Bangla (Sylheti)
- Mangalore Konkani
- Subzapuri

ii *Dravidian*
- Kannada
- Malayalam
- Manda
- Tamil
- Telugu
- Toda

iii *Austro-Asiatic*

iiia *Mon-Khmer*
- Khasi
- Rymbai
- Pnar

iiib *Munda*
- Ho
- Juang
- Kharia
- Mundari
- Santali
- Sora/Savara

iv *Tibeto-Burman*
- Ao (Mongsen)
- Bodo
- Dimasa
- Dumi
- Garo
- Hmar
- Kham
- Kokborok
- Konyak
- Ladakhi
- Manipuri
- Mao
- Mising
- Mizo
- Paite
- Rabha
- Sangtam
- Sema
- Tenyidie
- Thadou
- Tiwa

Most of the data on Tibeto-Burman languages and on some Munda languages were collected in field trips in the northeastern parts of India and in the states of West Bengal and Jharkhand.

It should also be mentioned that there are very few syntactic descriptions available for Tibeto-Burman, Munda and Mon-Khmer languages, especially at the level of detail and comprehensiveness which are necessary for useful analysis. Traditional grammars that are available do not address the conceptual issues that the linguist of today is interested in, in terms of the Universal Grammar and parametric variation.

The data collected and analyzed by the author with a team of researchers in the University Grants Commission's projects on syntactic typology, over a ten-year period, were also used in this work.