Language Universals and Linguistic Typology

SYNTAX AND MORPHOLOGY

Bernard Comrie

Second Edition

Basil Blackwell

Language Universals and Linguistic Typology

SYNTAX AND MORPHOLOGY

Bernard Comrie

Second Edition

Basil Blackwell

First published 1981
Reprinted with corrections and additions 1983
Reprinted 1986, 1987
Second edition 1989

Basil Blackwell Ltd
108 Cowley Road, Oxford OX4 1JF, England

British Library Cataloguing in Publication Data

Comrie, Bernard
Language universals and linguistic typology.
I. Grammar, Comparative and general—Word formation
I. Title
415 PZ41

ISBN 0-631-12971-5
ISBN 0-631-12618-X Pbk

Printed in Great Britain by
Billing & Sons Ltd, Worcester

CONTENTS

PREFACE TO THE SECOND EDITION

The continuing demand for *Language Universals and Linguistic Typology* necessitates yet another reprinting. In earlier reprintings, I restricted changes to corrections of typographical and other minor errors and updating of bibliographical information. However, developments since 1980 (when I completed the manuscript of the first edition), both specifically in universals and typology and more generally in grammatical theory, have rendered a somewhat more substantial reworking of the text essential. The main changes are in chapter 1, where a now outdated comparison of different data bases has been replaced by new material, and in chapter 11, which is substantially new. The other chapters remain more or less the same, though new material has been added to chapter 2; the main differences are clarifications of certain parts of the texts and a more substantial updating of the bibliography.

I am grateful to those who reviewed the first edition and to all those who have given me comments in various ways, not least among them my own students at the University of Southern California.

The basic aim of the book remains, as in the first edition, to provide an introduction to the study of language universals and typology from an, albeit slightly modified, Greenbergian perspective. I have resisted the attempt to engage in protracted dialogue with proponents of other approaches or even to provide extension commentary on such other approaches, as this would be out of place in an introductory text of this kind. I have retained essentially the same coverage of topics, though it should be realized that my selection of topics to illustrate work on universals and typology is in large measure personal.

It is perhaps worth emphasizing the main features of my outlook on language universals and typology as represented in the body of this book.

The languages of the world provide us with a rich variety of data on the basis of which we can study the general properties of the human language potential. (The richness of this material should not blind us to its fragility; languages are dying out at an alarming rate, a fact which raises suprisingly little public concern, even among those who do express alarm at the destruction of other parts of our environment.) In order to understand the human linguistic potential, we must develop methods, descriptions, and analyses that are not only sufficiently constrained to say interesting things about those properties that are common to all languages but also sufficiently flexible to permit insightful characterization of the degree of variation that we find among languages. Comparison of languages should be driven primarily by the varied data that languages present to us; while any comparison, or indeed any description, requires some degree of abstractness of representation, reliance on overly abstract approaches to linguistic description substantially increases the possibility that what will be compared are not languages but linguists' conceptions (or misconceptions) of languages. This does not imply an atheoretical stance, though it does imply an attitude that is at least cautious to some of the claims of currently dominant grammatical theories. Linguists working on universals and typology from the perspective advocated here are obliged to seek explanations for the cross-linguistic generalizations they establish; adoption a priori of a particular explanation (especially a largely untestable one such as innateness) serves only to blind the linguist to the possibility of alternative explanations. The approach that I present in this book is thus part of an attempt to provide an explanatory account of the nature of human language.

Los Angeles, August 1988 Bernard Comrie

PREFACE TO THE FIRST EDITION

After a period when the frontiers of linguistic research seemed to be concerned primarily with the analysis of English syntax, the last decade has seen a remarkable upsurge of interest in problems of language universals and linguistic typology using data from a wide range of languages. Despite the vast amount of work that has been carried out within this framework, there has been, to date, no general introductory work that has attempted to synthesize the main characteristics of this approach for the student of linguistics, who has had to turn almost from the very beginning to specialist literature on individual topics in article form. This book aims to fill this gap, to provide the advanced undergraduate and graduate student with an overview of the major current approaches to language universals and typology, with illustrations of the successes of this method – and also warnings about some of the dangers.

In a field where so much literature has arisen in a relatively short period, this book is necessarily very selective in the range of topics chosen, with preference for going into certain topics in depth rather than giving a superficial overview of the whole field. I have also restricted coverage, for the most part, to recent work on universals and typology, rather than try to give a historical account of earlier work in this area, although earlier work is mentioned, especially to the extent that it has not been subsumed by more recent research. Some of the selectivity necessarily reflects my own biases, towards those areas where I have worked myself or where I feel the most exciting results have been forthcoming. The book is concerned almost entirely with syntactico-semantic universals, although on occasion phonological universals are also used as illustrative material. I believe that critical discussion of work in a few areas is more valuable than an unannotated listing, however comprehensive, of claims that have been made about universals and typology.

The first two chapters are general in nature, presenting and arguing my view that the study of language universals can proceed most fruitfully on the basis of consideration of data from a wide range of languages, and embedding the study of syntactico-semantic universals within an integrative approach to language in which explanations for universals are sought not so much within the formal properties of language, but rather by relating formal properties of language, at various levels (including syntactic and phonological), to the extra-linguistic context within which language functions. Later chapters, for the most part, look at individual construction types or other syntactic phenomena, such as word order, relative clauses, causative constructions, case marking, from the viewpoint of universals and typological research that uses data from a wide range of languages within an integrative approach. The particular choice of topics is to a large extent arbitrary, reflecting my own interests, but if this choice is no better than some others, I would argue that it is also no worse.

It is difficult for me to give a comprehensive list of acknowledgements to all those who have contributed to the development of this book and the ideas contained in it: study of universals and typology necessarily interrelates with work on just about every other aspect of language and linguistics, and I must with regret refrain from a list of everyone who has influenced my thinking on language. The following acknowledgements therefore relate to those, in addition to linguists acknowledged specifically in the notes to the various chapters, who have influenced my thinking on universals and typology and who have influenced the particular mode of presentation adopted in this book.

My debt to Joseph H. Greenberg (Stanford University) will be apparent on almost every page: it is he, more than any other single linguist, who initiated the present interest in working on language universals on the basis of a wide range of languages, and who persisted in advocating this approach even in periods when it was far from fashionable. Edward L. Keenan (University of California at Los Angeles) helped me to see that interest in a wide range of languages is not incompatible with interests in theoretical and formal issues. My colleagues in the Department of Linguistics at the University of Southern California, sensing early on my conversion to an integrative approach to language in context, have provided an environment full of stimuli to the development of this interest.

Although at times I am necessarily critical, in this book, of the approach to language universals adopted within mainstream transformational-generative grammar, and especially by Noam Chomsky, I cannot and would not want to deny my indebtedness to my training within this model and to those who trained me in it. Whatever disagreements I have since developed with some of the tenets of the descriptive model and its ideological underpinnings, it has clearly raised syntactic analysis to a level of rigour and insightfulness without

which this book would not have been possible. Similar remarks apply to the model of syntax proposed by relational grammar: although I disagree with the emphasis on structure-internal explanations of syntactic generalizations, and on many other specific and general issues, this approach to syntax has provided me with a vast number of insights into syntactic structure that would otherwise probably have escaped me, and it is with genuine, not damning, praise that I would acknowledge that, as far as formal models of syntax go, relational grammar seems to me to go the furthest.

I have also benefited considerably from discussions with linguists at various institutions engaged in research into language universals and typology, to whom I have been able to present parts of my own work and who have in turn presented some of their work to me. In particular, I would mention the participants in the Linguistic Society of America Linguistic Institute at the State University of New York at Oswego (1976), with typology as one of its foci; the Stanford Universals Project; the Universals Project (Universalien-Projekt) of the Department of Linguistics of the University of Cologne; and the Structural Typology Group of the Leningrad Section of the Linguistics Institute of the Academy of Sciences of the USSR.

The materials contained in this book derive in large part from materials tested out on students who attended my courses and seminars on language universals and typology. I would therefore like also to thank all the students, faculty members, and visitors who attended these courses at the University of Cambridge, the Linguistic Society of America Linguistic Institute at the University of Illinois at Urbana-Champaign (1978), the University of Southern California, and the Australian National University. This book has, in addition, benefited from the comments of anonymous readers for Basil Blackwell and the University of Chicago Press.

Finally, I wish to express my gratitude to the many fieldworkers and native-speaker linguists, often working well away from the beaten track and well away from mainstream theoretical linguistics, who have both provided me with invaluable material for my work and encouraged me in this work by their interest in it and the possibility of a constructive dialogue between us. I hope they realize that my aim has not been to steal a relative clause or a causative construction from their language, but rather to put into practice my belief that the maximum benefit both to general linguistics and to the description of individual languages will develop from the maximum integration of these two approaches – the one cannot flourish without the other. Or more generally: linguistics is about languages; and languages are spoken by people.

Los Angeles, January 1981 Bernard Comrie

PREFACE TO THE SECOND PRINTING

This second printing has given me the opportunity to update bibliographic references and to correct some misprints and minor errors of fact (fortunately, not affecting the points illustrated), as well as to improve on some unclear formulations. In addition to points made by reviewers, I am grateful for comments from Winfried Boeder, Peter Cole, R. M. W. Dixon, Andrew Goodson, Kim Jong-mee, Herbert H. Paper, William Rutherford, and Sandra A. Thompson.

Los Angeles, October 1982 Bernard Comrie

1

LANGUAGE UNIVERSALS

In this chapter, a number of general issues relating to the study of language universals are considered, and a particular approach to language universals research is advocated, in contrast to other possibilities. The exemplification of general points necessarily refers to individual topics that are discussed in the body of the book. The reader unfamiliar with the background to the relevant issues may find several parts of chapter 1 unclear at first reading; such readers are advised initially to read through chapter 1 relatively quickly, returning to more detailed study of its claims after they have familiarized themselves with the body of the book.

1.1 APPROACHES TO LANGUAGE UNIVERSALS

1.1.1 TWO MAJOR APPROACHES

In this section, we will contrast two major methodological approaches to language universals that have been adopted in recent linguisitic work. The two approaches can be contrasted on a number of parameters, the most important of these being the following: the data base for research on language universals (a wide range of languages, or a highly restricted set of languages); the degree of abstractness of analysis that is required in order to state language universals (for instance, in terms of more concrete or more abstract syntactic analysis); and the kinds of explanations advanced for the existence of language universals. The individual parameters, and others, will be taken up again in subsequent sections. Although each of these parameters is logically independent of the others, in fact the two major recent approahces each represent a coherent clustering of these parameters. On the one hand, some linguists have argued that in order to carry out

research on language universals, it is necessary to have data from a wide range of languages; linguists advocating this approach have tended to concentrate on univerals statable in terms of relatively concrete rather than very abstract analyses, and have tended to be open, or at least eclectic, in the kinds of explanations that may be advanced for the existence of language universals. On the other hand, some linguists have argued that the best way to learn about language universals is by the detailed study of a small number of languages; such linguists have also advocated stating language universals in terms of abstract structures and have tended to favour innateness as the explanation for language universals. The first of these two approaches is perhaps most closely associated with the work of Joseph H. Greenberg and those inspired by his work, and also reflects the orientation of the present book. The second is most closely associated with the work of Noam Chomsky and those most directly influenced by him, and might be regarded as the orthodox generative position.

At first sight, at least with regard to the data base for work on language universals, it might seem that Greenberg's approach is necessarily correct, since surely in order to establish that something is universal in language one would need to look at a wide range of languages – if not, indeed, at all languages. However, the argumentation is by no means so simple as this, a point to which we will return in section 1.1.2. For the remainder of this section, we will outline the motivation for adopting Chomsky's approach to language universals. Although this argumentation is, as the subsequent discussion will show, vulnerable on a number of points, both conceptual and empirical, it does represent a coherent position with regard to language universals research which cannot simply be ignored.

A generative description of a language, or more specifically of the syntax of a language (although similar arguments could be transferred, say, to a generative phonological description), maintains that syntactic representations are highly abstract objects, considerably removed from anything observable in the linguistic data. The abstractness of syntactic structures, or at least some levels of syntactic representations, characterizes most versions of generative grammar, including in particular government and binding.

When the existence of such abstract representations is taken into account in discussing the way in which children acquire their first language, a potential problem arises. If the best way of characterizing the structure of a language involves abstract structures, then it is probably justifiable to assume that, in acquiring a language, the child internalizes these abstract structures. This implies, in turn, that he must also internalize rules for passing from these abstract structures to the more concrete levels of analysis. The argument then continues by claiming that there is no way, in

terms of our current knowledge of learning abilities, in which the child, presented only with the data of adults using the language around him, could induce these abstract principles from these data. Moreover, it is argued that the rules needed to pass from abstract to concrete structure are subject to a number of highly specific constraints: again, it is not clear how these constraints, which are themselves highly abstract in nature, could be induced by the child from the raw data presented to him by adult speech. More generally, if the child is viewed simply as a tabula rasa, as having no predisposition to analysing data in terms of one formal system rather than any other one, then it is difficult or impossible to explain how the child does in fact come to acquire his first language within a relatively short period of time.

This learnability problem evaporates if one makes the crucial assumption underlying orthodox generative work on language univerals. The reason why the child acquires his first language so effortlessly is that the crucial abstract principles of generative grammar are innate: they are available to the child from birth (or, perhaps, are available from a certain period soon after birth as part of the maturnational process, but at any rate are preprogrammed at birth), so that the child does not have to learn them, but can use them in figuring out which particular language, of those permitted by the general theory of generative grammar, is being spoken in his speech community: although the general theory (and, equivalently, the set of innate abstract principles internal to the child) allows an infinite number of possible languages, the types of languages are greatly restricted to those permitted by the constraints imposed by the theory.

Given the simple observation that children learn their first language so readily, one might wonder whether an even stronger claim could not be made, namely that the language as a whole is innate. This would assume that a child born into a given speech community is already preprogrammed with knowledge of the language of the speech community, presumably having inherited it from his parents. However, further observation soon shows that this scenario, though clearly simplifying the learnability problem, cannot be correct. It would imply that a child could only learn, or at least would much more readily learn, the language of his parents, irrespective of the language of the surrounding community. Now, it is known from observation that children acquire, with approximately equal facility, the language of whatever speech community they happen to grow up in, quite irrespective of the language of their parents or their more remote ancestors; this can be seen most clearly in the case of children who are brought up by speakers of a language different from that of their natural parents. Thus it cannot be the case that the language as a whole is innate – note that this was established on the basis of empirical observation, rather than by speculation. At best certain

principles common to all human languages would be innate, which would thus facilitate the child's task in acquiring whichever language he happens to be exposed to, with no preference for one language over any other. This now brings in the last link in this argument: since the abstract principles claimed to be innate are the same for all children, irrespective of ethnic background, they must be neutral with respect to differences among languages, i.e. they must be universal. One can thus establish an equation between language universals and innate ideas: language universals would be those innate linguistic principles which facilitate the child's language-learning task.

In fact, the position is slightly more complex than this when one looks at the way in which mainstream generative grammar has developed over the last decade. In addition to innate principles that are common to all languages, generative grammarians argue that there is also an innate set of parameters. Each parameter has a number of possible settings, and languages can differ in the setting of a particular parameter; we return to this in section 2.4. The possibility of different languages having different parameter settings, so-called parametric variation, would then account for the systematic typological variation among languages. It is, of course, an empirical question how much typological variation there is among languages, and generative grammar typically follows the line that such variation is likely to be highly restricted. Although it is, within this approach, necessary to examine a typologically diverse set of languages in order to uncover the innate principles of parametric variation, the approach still highly favours the detailed abstract study of a small number of languages rather than, as advocated by those who follow the Greenbergian paradigm, the study of a broader range of languages as a prerequisite to gaining an understanding of cross-linguistic typological variation.

Once one makes this equation, it is but a short step to justifying the methodology of language universals research adopted by Chomsky. Since the universals in which one is interested are abstract principles, there is no way in which the analysis of concrete data from a wide range of languages would provide any relevant information. Rather, one should investigate relations between abstract and more concrete levels of representation, in order to factor out the abstract principles which constrain language structure (and which are, thus, language universals or, equivalently, innate ideas). In terms of the weighting of breadth as against depth of coverage of languages, the position outlined in this section clearly favours concentrating on depth, with preference for the detailed abstract study of a small number of languages rather than casting the net more widely but with less depth. From this come the general methodological tenets outlined at the beginning of the section: the most profitable way to study language universals is to study a small number of languages in depth, in terms of an

abstract analysis of those languages – the universals themselves are then of an abstract nature (abstract constraints on a system involving abstract levels of representation); since language universals are equated with innate ideas, the latter provide an obvious explanation for the former, and the only way in which one might need to extend the consideration of explanatory principles would be to ask whether there is in turn an explanation for the innate ideas.

In sections 1.1.2–3, we will consider practical reasons why this research strategy for language universals, despite its internal coherence, suffers from a number of serious defects, leading to its rejection in the present book. To conclude this section, however, we will examine some more general weaknesses of the argumentation leading to the research paradigm. These weaknesses stem mainly from the fact that the argumentation is almost entirely aprioristic, with virtually no appeal to actual data supporting the position being argued for: indeed the only direct appeal to facts, namely that children learn any language with comparable facility, served only to establish a non-universal (the specific language as a whole cannot be innate). Of course, in any science it is necessary to establish hypotheses which may, in the initial stages, be largely aprioristic, but it is important then to test out these hypotheses, to see to what extent they do fit with the data range to be explained. The real problem with the kind of aprioristic argumentation summarized in this section is that it is not, given present techniques, subject to any kind of empirical test, i.e. is not potentially disconfirmable. More specifically, the claims about what is inherently easy or difficult to learn are not based on any actual research on ease of learnability, so that again one must simply take on trust that some things are easily learned and others less so, others perhaps being impossible to learn. Finally, as will be shown in more detail in section 1.1.3, any argument based on an abstract analysis is no stronger than is that abstract analysis itself, and given the wide range of competing abstract analyses of, say English syntax, one must again simply take on trust that one analysis, rather than another, is the psychologically real analysis (or, at least, the best that we can, in our present state of knowledge, advance as the psychologically correct analysis). Generalizing these remarks, the research paradigm outlined in this section is characterized by a number of questionable assumptions that are crucial to the argumentation, these assumptions being for the most part untestable, at least at present, so that acceptance of the paradigm becomes simply a matter of faith.

1.1.2 THE DATA BASE

In this section, we will establish some more practical reasons why the study of language universals must operate with data from a wide range of

languages, then look at some of the implications of this for the practice of research on language universals. A priori, there seems to be no reason to assume either that language universals research should require a wide range of languages or that it should not, and one can easily come up with analogies from other investigations for either of these two positions. For instance, if one wanted to study the chemical properties of iron, then presumably one would concentrate on analysing a single sample of iron, rather than on analysing vast numbers of pieces of iron, still less attempting to obtain a representative sample of the world's iron. This simply reflects our knowledge (based, presumably, on experience) that all instances of a given substance are homogeneous with respect to their chemical properties. On the other hand, if one wanted to study human behaviour under stress, then presumably one would not concentrate on analysing the behaviour of just a single individual, since we know from experience that different people behave differently under similar conditions of stress, i.e. if one wanted to make generalizations about over-all tendencies in human behaviour under stress it would be necessary to work with a representative sample of individuals (even if the study were restricted to a single society, let alone if one envisaged a cross-cultural study).

Since one of the things we want to find out in work on language universals is the range of variation found across languages and the limits placed on this variation, it would be a serious methodological error to build into our research programme aprioristic assumptions about the range of variation. Moreover, as we shall see in the following paragraphs, there is evidence from fairly basic research on language universals that, in certain crucial cases that have arisen in work to date, data from a wide range of languages were in fact necessary to have a reasonable chance of validating a given language universal.

First, there are certain language universals that simply cannot be predicated of an individual language. In particular, implicational universals are of this kind. We shall look at implicational universals in more detail in section 1.2.2, for the moment it suffices to note that an implicational universal always involves at least two linguistic properties, which we may symbolize as *p* and *q*, related to one another as an implication (condition), i.e. '*if p, then q*'. As a simple example, we may take the following actual example: if a language has distinct reflexive pronouns (i.e. distinct from non-reflexive pronouns) in the first or second person, then it has distinct reflexive pronouns in the third person. In this example, the property *p* is 'having distinct first or second person reflexive pronouns' and g is 'having distinct third person reflexives'. The combination of these two properties can be seen, for instance, in English, which has *I hit myself* and *he hit himself*. Note, however, that English does not provide evidence for stating the universal as

an implication: indeed, if we were only investigating English, we might be led to conclude that a language must have distinct reflexive pronouns in all persons. Investigation of other languages, however, soon shows this generalization to be untrue. French, for instance, has distinct reflexive pronouns in the third person but not in the first or second person, as in *je me frappai* (cf. *il me frappa* 'he hit me') but *il se frappa* (cf. *je le frappai* 'I hit him'). Anglo-Saxon has no distinct reflexive pronouns in any person, as in *ic slog me* (cf. *he slog me*) and *he slog hine*, which latter can mean either 'he hit him' (two different people involved) or 'he hit himself'. The fourth logical possibility – a language having distinct reflexive pronouns in the first or second person but not in the third person – is excluded by the implicational universal.

If we were to base our study on any single language, then we would be led to make a statement stronger than the implicational universal, as we noted above with respect to the English data. Only consideration of data from a range of languages enables us to see that of the four logical possibilities – (a) distinct reflexive pronouns in first/second person and in third person, (b) distinct reflexive pronouns in first/second person but not in third person, (c) distinct reflexive pronouns in third person but not in first/second person, (d) distinct reflexive pronouns in neither first/second nor third person – one, namely (b), is systematically absent.

Of course, each individual language must be consistent with an implicational universal, otherwise it would be a counterexample, but no individual language provides the kind of evidence that would be needed to justify positing an implicational universal. (The only exception to this would be where a given individual language has more than one construction in a given area, in which case it might be possible to establish an implication on the basis of data from the two constructions within the same language; this possibility is illustrated in chapter 7, for relative clauses.)

In addition to such examples where data from a range of languages is absolutely necessary, even aprioristically, in the establishment of language universals, there are other examples where failure to consider a range of languages has led to the positing of putative language universals which then crumble as soon as one is presented with data from other languages. As illustration, we shall take one example from $\bar{X}$-theory, part of the extended standard theory of generative grammar. Here, it is argued that, if we treat $\bar{X}$ as a cover symbol for various kinds of phrase (e.g. noun phrases, verb phrases, adjective phrases), then there is a general (i.e. language-independent) expansion rule $\bar{X} \rightarrow \text{Spec}_{\bar{X}}\ \bar{X}$, i.e. that a phrase $\bar{X}$ (a phrase whose head constituent is X, so that noun phrase would be symbolized $\bar{N}$) consists of the immediate constituents Specifier-of-$\bar{X}$ and $\bar{X}$. In terms of actual phrase types, if X is noun, then $\text{Spec}_{\bar{X}}$ would be, for instance, an

article (determiner); if X is verb, then $Spec_{\bar{X}}$ would be an auxiliary. The language-independent schema given above for the expansion of $\bar{X}$ says nothing about the relative order of $Spec_{\bar{X}}$ and $\bar{X}$, this being left as a parameter on which individual languages may vary. However, the schema does make claims about the relative order of $Spec_{\bar{X}}$ and $\bar{X}$ across phrase types within a given language. Interpreted as an absolute, exceptionless universal (see section 1.2.3), it claims that in a given language, for all phrase types either the Specifier precedes or it follows, i.e. either determiners precede nouns and auxiliaries precede verbs, or determiners follow nouns and auxiliaries follow verbs. Interpreted as a tendency, it says that languages would tend to adhere to this generalization, although it would always be open to an individual language to violate the universal.

This universal was originally proposed on the basis of English data, and in English it is indeed the case that determiners precede their noun (as in *the book*) and that auxiliaries precede their verb (as in *must go*). There are clear counterexamples to the principle as an absolute universal: for instance, in Malay determiners follow their noun (e.g. *surat itu* 'that letter', literally 'letter that'), while auxiliaries precede their verb (e.g. *sedang membaca* 'is reading', *akan membaca* 'will read'). In fact, in current work within the extended standard theory, it is usually stated, or at least allowed as a possibility, that the schema may be a tendency, rather than an absolute universal. However, even this claim turns out to be invalid as an attempt to characterize variation across languages. The number of languages in which determiners follow nouns and auxiliaries follow verbs is small, while there are many languages – including most languages of the widespread canonical SOV type (see chapter 4) – that have determiners preceding the noun but auxiliaries following the verb, as in Japanese *kono hon* 'this book', *aisite iru* 'loves', literally 'loving is'. In other words, in terms of the actual distribution of word order types along these two parameters (determiner relative to noun, auxiliary relative to main verb), the schema makes incorrect claims, even as a statement of a tendency.

Note that the weakness of the one-language approach to language universals illustrated here is not simply that a given putative universal turns out to be wrong. This is almost inevitably going to be the case whatever data base one adopts for research on language universals, since certain attested language types are simply very rare, and might very well not be included even within a comprehensive sample of the world's languages: for instance, click consonants as regular phonemes are restricted to Khoisan and neighbouring Bantu languages in southern Africa; very few languages, perhaps restricted to the Amazon basin, have object – verb – subject (OVS) as their basic word order. The point is rather that, once the putative universal concerning the order of $Spec_{\bar{X}}$ and $\bar{X}$ was formulated it was

necessary to establish, say by looking even at a few other languages with different word order possibilities, whether or not their generalization stood some chance of being a valid cross-language generalization.

From the theoretical and practical deficiencies of trying to work on language universals on the basis of a single language, one might think that the ideal would be to base the study of language universals on simultaneous investigation of all languages of the world. However, there are two very obvious reasons why this is impossible. First, we know that many languages have become extinct without ever having been recorded, or without being recorded in sufficient detail to be of value to our enterprise, in addition to which, given language change, many new languages will arise in the future; clearly, these two sets of languages are unavailable to us, and therefore a large number of actual human languages (defined as languages that were, are, or will be spoken) are not amenable to investigation. Secondly, the estimated number of languages spoken in the world today is so large that, if we were to await investigation of each language before embarking on research on universals, we would probably never even initiate our main task. Although estimates of the number of different languages vary considerably, they tend to cluster around the 4,000 mark.

In practical terms, then, the problem with which we are faced is establishing a representative sample of human languages in order to be able to carry out work on language universals that is both manageable in practical terms and likely to be free of bias from concentrating unduly on a single language or group of languages. The population from which we draw our sample is limited to the languages actually spoken today, plus some of the better documented extinct languages (though, given the absence of native speakers, certain questions concerning a dead language will necessarily go unanswered). Behind this statement, there are two assumptions that are necessary to such work on language universals, but which should not go unstated. One is that, at least within a time-span of several thousand years in either direction from the present, there has been no significant sense in which human language has evolved, i.e. no sense in which human language as a whole today is different in essence from that of ten thousand years ago; more specifically, it assumes that all human languages spoken today represent the same level of evolution. The more specific assumption seems reasonable, given that no structural features of language have been found that can unequivocally be correlated with more or less civilized social structure (however the latter is defined). The more general assumption is, however, beyond the possibility of empirical confirmation or disconfirmation, but lies at the basis of all work, of whatever orientation, that treats human language as a homogeneous phenomenon.

The second assumption is that the range of human languages spoken in

the world today is sufficiently large and varied to include examples of virtually all the kinds of structure that are possible in human language. This second assumption is much more questionable than the first, especially since we know that in certain respects the languages of the world are decidedly skewed in favour of certain structures and against others: thus, click consonants are restricted to a small part of southern Africa, languages with object-initial word order seem restricted to one part of South America, whereas languages with verb-final word order can be found in every continent. What, then, if this assumption should turn out to be false? In this eventuality, research on language universals will simply be impossible, and in practice researchers working within the same paradigm of language-universals research as is presented in this book simply make the assumption that the range of attested languages is sufficiently wide. In practice, significant results have been achieved by making this assumption. Moreover, although there is skewing on certain parameters, there are other parameters (for instance in the syntax of relative clauses; see chapter 7) where representatives of different types are found scattered across the world, so that in certain areas, at least, we can be reasonably sure that the totality of the world's languages does represent a reasonable population from which to draw our sample – it is not just the case that this is the only population we have.

Assuming that we have a reasonable population, the next problem is to decide on what kind of sample we are going to use from this population, given the impracticality of attempting to work with all the world's languages. Here, there are certain biases that must be avoided in establishing the sample, although not all work on language universals has necessarily done so. First, it is essential that the languages chosen in the sample must be from a range of genetic language families. Since members of a single language family, by definition, have certain traits in common because they have inherited them from their ancestor language, restriction of the sample to a single language family would not enable us to distinguish between common properties that are genuine language universals and those that are chance properties of the given genetic group. Likewise, biasing the sample in favour of one language family would give the impression that accidental structural properties common to that language family are in fact more widespread than they are. With respect to guarding against genetic bias, there is a specific proposal in the literature, devised by Alan Bell, which we shall outline briefly.

Bell argues that, in establishing a sample of languages, one should ensure that each 'group' of languages should be given equal representation, where a group is defined as a genetically related set of languages separated from their common ancestor by a time-depth of 3,500 years. On

this criterion, for instance, the Indo-European family would consist of 12 groups. Bell gives the following as the number of groups in the language stocks of the world:

Dravidian	1	Na-Dene	4
Eurasiatic	13	Austric	55 (approximate)
Indo-European	12	Indo-Pacific	100 (estimate)
Nilo-Saharan	18	Australian	27 (approximate)
Niger-Kordofanian	44	Sino-Tibetan	20 (approximate)
Afroasiatic	23	Ibero-Caucasian	4
Khoisan	5	Ket	1
Amerind	150 (estimate)	Burushaski	1

(Many of the stocks are dubious as established genetic units – for instance, Amerind, which groups together virtually all the native languages of the Americas, or Indo-Pacific, which groups together all the non-Austronesian languages of New Guinea – but in terms of the ratio of languages from different stocks, whether or not individual stocks are considered to be established genetic units, this table does provide a reasonable working hypothesis.) The total number of groups is 478, so that in a sample of 478 languages each stock should be represented by one language from each group. In practice, any sample will almost certainly be smaller, to achieve a reasonable practical compromise between depth and breadth of coverage, although the ratios will of course be the same. With a smaller sample, one disadvantage is that stocks with only one or a few groups tend automatically to be excluded. In terms of work to date on language universals, it is clear that the samples used involve many skewings, which, while not unexpected, does seriously call into question whether or not their results are representative of human language as a whole. Thus Indo-European languages are grossly overrepresented, for obvious social reasons: speakers of these languages are more readily available, and grammars of these languages are more readily available. At the other extreme, languages of New Guinea, which ought to make up about 20 per cent of a representative sample, are usually completely missing, as are Amazonian languages: speakers of the relevant languages are rarely available outside New Guinea or South America, and there are few detailed grammars of any of these languages. Until good descriptions of a wider range of languages are available to linguists working on language universals, this skewing is likely to remain in their samples, even where the existence of the skewing and of its disadvantages are recognized.

In addition to guarding against genetic bias, it is also necessary to guard against areal bias, i.e. against selecting an unrepresentatively large number of languages, even if from different genetic groups, from the same geographical

area. This is because, as discussed in more detail in section 10.2, languages spoken in the same geographical area tend over time to influence one another and come, through borrowing or shared innovation, to have features in common that are not necessarily language universals, or even particularly frequent cross-linguistically. A good example would be the diffusion of click consonants from the Khoisan languages into neighbouring Bantu languages. Therefore, in addition to ensuring that the languages in a sample are representative genetically, they must also be representative areally. It has recently been suggested that, given the extent to which languages have influenced one another by contact, it may actually be impossible to meet the constraint against areal bias fully, since any sample that is large enough to be genetically representative will necessarily also include languages that have been in areal contact with one another. But at least in establishing a sample free of genetic bias, one should also choose the individual languages so that, as far as possible, no two languages are picked that are known to have been in close areal contact.

In addition to these two obvious biases that should be guarded against, and against which it is relatively easy to guard (at least in theory, i.e. not taking into account practical problems of availability of language material), in an ideal sample one would also want to guard against biases in favour of or to the detriment of classes of languages defined by major typological features. For instance, it would be quite possible to come up with a sample of languages that would be representative genetically and areally, but where all the languages, or at least an overwhelming majority, would have the basic order subject – object – verb, this being the most frequent basic word order in the world's languages. In particular, where it is known, hypothesized, or suspected that a given typological variable may correlate with the phenomenon under investigation, care should be taken to guard against such typological bias.

In sum, then, to carry out detailed work on language universals one needs a representative sample of languages, representativeness being defined in particular as absence of genetic, areal, or typological bias.

1.1.3 DEGREES OF ABSTRACTNESS

In section 1.1.1, we noted that one of the differences between the two main approaches current in research on language universals concerns the degree of abstractness that is involved in stating language universals. Within Chomsky's approach, language universals are primarily constraints on the relation between abstract structures and more concrete structures, i.e. necessarily involve a considerable degree of abstractness. In Greenberg's approach, on the other hand, universals are stated primarily in terms of more concrete

levels of analysis. In this section, we shall illustrate these differences in somewhat more detail, concentrating in particular on the following two questions. The first is whether there is any validity to surface structure universals, i.e. to universals that require only a minimum of abstract analysis. The second will be the empirical status of universals that require reference to very abstract analyses. Throughout the discussion, it is important to bear in mind that we are not dealing with a strict dichotomy between abstract statements on the one hand and concrete statements on the other, but rather that there is a continuum between the two. Thus, many of the specific universals that have been proposed by Greenberg and those influenced by him require some degree of abstractness. Greenberg's original contribution on word order typology, by referring to such parameters as the relative order of subject, verb, and object in the clause, assumes that it is possible to identify the subject of an arbitrary clause in an arbitrary language. However, identification of a subject requires a certain amount of abstract analysis (there is no physical property that is common and unique to subjects across all sentences of all languages), and indeed, as we shall see in chapter 5, there is considerable controversy surrounding the identification of subject in many sentence types in many languages, and even concerning the validity of the notion subject at all. But it does still remain true that a statement about the nature of surface-structure subjects is less abstract than one about underlying subjects.

The answer to the first question posed above, namely whether or not there are any valid concrete universals, is in a sense given illustratively by the discussion of the body of this book, concerned as it is with a range of recent proposals concerning surface structure universals. Moreover, in discussing not only proposals for actual language universals but also suggested explanations for language universals in these later chapters, it should become clear that concrete universals can not only be established with a degree of rigour that is not possible with more abstract formulation, but that such universals can then be integrated into a much broader perspective on human language than is possible with purely formally stated universals, irrespective of the degree of abstractness required in their formulation.

In the present section, we shall concentrate, therefore, on the second of the two questions posed above, the empirical validity of abstract universals, the crucial point here being that an abstract universal is no stronger than (and may even be weaker than) the analysis on which it is based, i.e. if a given abstract analysis is controversial then so too will be any universal that builds on it. Rather than giving a general discussion of the pros and cons here, we shall examine a specific example that has arisen in recent work in relational grammar, an offshoot of transformational grammar which, while rejecting some of the tenets of transformational grammar, does share with it a predi-

lection for stating universals in terms of abstract structures. A number of the world's languages have a so-called impersonal passive construction, in which, in surface structure, the verb has no overt subject, the agent being expressed, if at all, as an agentive phrase; however, the objects of the verb, including the direct object of a transitive verb, remain just as in the ordinary active sentence. This can be contrasted with the English-type (personal) passive, where there is an overt subject, corresponding to the direct object of the active sentence. We may illustrate the impersonal passive by some examples from Welsh:

Lladdodd y ddraig y dyn. (1)
killed the dragon the man
'The dragon killed the man.'

Lladdwyd y dyn gan y ddraig. (2)
killed-PASSIVE the man by the dragon
'The man was killed by the dragon.'

In stating the relationship between the active and passive sentences, a straightforward, superficial statement would be to say that the subject of the active corresponds to the agentive phrase of the passive, with the result that the passive has no overt subject. However, this violates two putative universals of relational grammar. According to the Motivated Chomage Law, it should be impossible for the subject of the active to be demoted to an agentive phrase unless some other noun phrase is advanced to subject position (i.e. the demotion of the subject is contingent upon the advancement of some other noun phrase into that position). The Final 1 Law says that a clause invariably has a final subject, i.e. a subject at the end of the operation of all cyclic rules. Sentence (2) clearly lacks a surface subject, and this is not at issue. In order to maintain the validity of these putative universals, it is necessary to assume, within relational grammar, that some noun phrase (a dummy noun phrase, whose origin is not at issue here) is inserted into subject position (more accurately: inserted into direct object position then advanced to subject) in the impersonal passive construction, thereby causing the demotion of the original subject; the dummy subject does not show in surface structure, or at least has no phonological realization.

We must now consider whether or not this putative universal involves any empirical claim. On the basis of the data and discussion given here, there is no empirical claim involved. If this analysis is available for impersonal passive constructions, then it is clearly impossible to construct a set of data that would be a counterexample to the Motivated Chomage Law and/or the Final 1 Law, since proponents of these laws would simply say that at an

intermediate level of abstractness the sentence in question does indeed have a subject, only this subject is never realized overtly. Note that we cannot say that the analysis proposed within relational grammar is wrong, in the sense that there are counterexamples to it; rather, this analysis makes no empirical claim, so that it is impossible to construct even a potential counterexample to the hypothesis.

In the present work, it will be maintained that the only language universals that are of empirical interest are those to which potential counterexamples can be constructed. Putative universals which simply test the ability of linguists to come up with abstract analyses that are consistent with any conceivable set of data may tell us something about linguists, but they do not tell us anything about language.

1.2 CLASSIFICATION OF LANGUAGE UNIVERSALS

1.2.1 FORMAL AND SUBSTANTIVE UNIVERSALS

In the generative literature on language universals, one distinction that has played a major role is that between formal and substantive universals. Although this distinction will play a smaller role in the present book, some discussion of the distinction will be necessary, if only to place the present work in its broader context.

Substantive universals are those categories, taken in a wide sense, that are posited as language universals. In syntax, for instance, they might include such categories as verb, noun, noun phrase, subject, direct object, main verb. In phonology, a clear example would be the distinctive features of Jakobsonian phonology. Although substantive universals delimit the class of possible human languages relative to the class of logically possible languages, they can do so in two ways. On the one hand, a substantive universal may be a category that must be present in each individual human language (in phonology, vowel would be a good candidate). On the other hand, the set of substantive universals in a given area might represent a set from which individual languages select a subset, i.e. they would define the total range available to natural languages, items from outside this range being defined as impossibilities. This second possibility is again well represented by the Jakobsonian theory of distinctive features, which claims that the phonological system of any arbitrary language will make use of no distinctive feature not contained in the list, although it is not necessary that any individual language should make use of the whole set (thus English does not make distinctive use of the feature Checked). Another way of characterizing the difference between the two types would be as follows:

the first distinguishes what is necessary in a language from what is unnecessary, the latter distinguishes what is possible from what is impossible. In combination, they distinguish necessary characteristics of human languages, possible characteristics, and impossible characteristics.

Formal universals are rather statements on the form of rules of grammar. Again, it would be necessary here to distinguish among necessary, possible, and impossible properties of rules of grammar in human language. As an example, we may take the claim that no language can have a formal rule that operates by giving a left–right inversion of a string of arbitrary length. In slightly more concrete terms, this says, for instance, that no language could form questions by simply inverting the word order, so that the question corresponding to *this is the house that Jack built* would be *built Jack that house the is this?* This particular formal universal seems to stand the test of verification in a wide range of languages; it is a special case of a more general formal universal proposed within transformational-generative grammar, namely that transformations are structure-dependent operations, to which we return in section 1.2.3.

Throughout most of the development of generative grammar, it has been held that the constraints delimiting the class of possible rules are formal universals, and indeed most of the work on universals within this approach to syntax has been concerned with just such formal constraints. However, there have also been suggestions that at least part of the problem of delimiting the set of rules might be in terms of substantive universals, such that there would be a certain set of rules, subject to variation in detail in individual languages, from which the individual language would select, at least in order to build up its core syntactic processes. One candidate for such a rule would be passive (personal passive), characterized as a process whereby the original subject is deleted or demoted to an agentive phrase while the original object is advanced to subject position; beyond this core, individual languages would vary, for instance, as to whether or how they mark the voice change on the verb or the noun phrases. Thus English uses the auxiliary *be* with the past participle to mark the voice change on the verb, and the preposition *by* to mark the agentive phrase in the passive, as in *the man was hit by the woman*, whereas Latin would use a different ending on the main verb, in addition to changes in the case of the noun phrases, e.g. active *mulier* (NOMINATIVE) *hominem* (ACCUSATIVE) *videt* 'the woman sees the man', passive *homo* (NOMINATIVE) *ā muliere* (preposition + ABLATIVE) *vidētur* 'the man is seen by the woman'. In the present work, the existence of such substantive universals plays a significant role, as can be seen from the treatment of such topics as the cross-language comparison of relative clause constructions (chapter 7).

1.2.2 IMPLICATIONAL AND NON-IMPLICATIONAL UNIVERSALS

For certain properties of language, it seems that we can state whether or not they are found in natural language without reference to any other properties of the given language. For instance, the statement that all languages have oral vowels makes absolutely no reference to any other items that must or must not be present. Such universals are non-implicational. Many other statements about language universals, however, relate the presence of one property to the presence of some other property, i.e. state that a given property must, or can only, be present if some other property is also present. In section 1.1.2, an example of an implicational universal was introduced: if a language has first/second reflexives, then it has third person reflexives. As illustrative material, we shall discuss this example more thoroughly. Two properties are involved: the presence or absence of first/second person reflexives, and the presence or absence of third person reflexives. Let us symbolize presence of first/second person reflexives as p (whence absence of first/second person reflexives is *not-p*), and the presence of third person reflexives as q (whence absence of third person reflexives is *not-q*). The universal can now be symbolized: if p, then q. Logically, there are four possibilities for combining these various parameters:

(a) p and q

(b) p and *not-q*

(c) *not-p* and q

(d) *not-p* and *not-q*

The implicational statement is to be interpreted (by definition) rigidly in terms of the interpretation of material implication in standard propositional calculus, which means that if the implicational statement '*if p, then q*' is true, then the possibilities (a), (c), and (d) above are allowed, whereas logical possibility (b) is disallowed. In section 1.1.2, we demonstrated that this is indeed the case with our particular example: there are languages like English with both first/second person reflexives and third person reflexives (type (a)); there are languages like French with no first/second person reflexives but with third person reflexives (type (c)); there are languages like Anglo-Saxon with no first/second person reflexives and no third person reflexives (type (d)); but type (b) – first/second person reflexives but no third person reflexives – is not attested. In formulating implicational universals, it is important that the rigid interpretation of material implication be followed,

and in particular to note that a given implicational universal always allows three of the logical possibilities while disallowing one; only attestation of the disallowed fourth logical possibility counts as a counterexample to an implicational universal.

Although it is important always to keep in mind the logical definition of implication, in order to avoid making pointless language universal statements there is one other factor that should be borne in mind, namely that in order for an implicational universal to be a reasonable claim to make, each of the three permitted possibilities should in fact be represented. As an example of a universal that falls foul of this requirement, we may note the following: if a language has nasalized vowels, then it also has oral vowels. In a sense, the universal is true, and certainly there are no counterexamples, i.e. no languages which have nasalized vowels (*p*) but lack oral vowels (*not-q*). However, of the three permitted possibilities, only two are in fact attested: languages with both nasalized and oral vowels (*p* and *q*), and languages with oral vowels but no nasalized vowels (*not-p* and *q*); there are no languages with no vowels at all (*not-p* and *not-q*). In a situation like this, where one of the classes has no representatives, one can in fact make a stronger claim, in this case the non-implicational universal: all languages have oral vowels. This, together with the statement that nasalized vowels are possible, renders the original implicational universal superfluous.

In addition, the most significant kind of implicational universal are those where there is a reasonably large number of languages within each of the three permitted classes. An obvious example of an implicational universal that fails to meet this criterion of significance would be the following: if a language is English, then the word for the canine quadruped is *dog*. Case (a) is represented by one and only one language, namely English; case (b), the excluded logical possibility, is indeed not represented, i.e. there are no counterexamples; case (c), i.e. a language which is not English but has the word *dog* in this meaning, has at least one member, the Australian language Mbabaram; case (d), i.e. languages which are not English and which have a different word for the canine quadruped, comprise probably all the other languages of the world. As illustration, an obviously stupid example was chosen here simply to illustrate the general point – presumably no-one would seriously have proposed this as a significant language universal; but it is important to guard against introducing the same deficiency in more covert form. For instance, if a given property or set of properties is only found in a single language in the sample, then any implicational statement that includes this property or set of properties as *p* may in fact merely be stating a property that is peculiar to that one language. In the present state of our knowledge of object-initial languages, for instance, with only one such language described in detail (the Carib language Hixkaryana), it would be premature to attempt to correlate its object-initial word order as *p* with any other properties as *q*.

1.2.3 ABSOLUTE UNIVERSALS AND TENDENCIES

Another parameter along which universals can be classified is that distinguishing absolute universals, i.e. those that are exceptionless, and those that exist as tendencies, but do still have exceptions. This distinction is independent of that between implicational and non-implicational universals, giving over all a fourfold classification. There are absolute non-implicational universals, such as: all languages have vowels. There are absolute implicational universals, such as: if a language has first/second person reflexives, then it has third person reflexives. There are non-implicational tendencies, such as: nearly all languages have nasal consonants (although some Salishan languages have no nasal consonants). Finally, there are implicational tendencies, such as: if a language has SOV basic word order, it will probably have postpositions (but Persian, for instance, is SOV with prepositions).

One question that immediately arises here is whether it is justifiable to talk about something being a language universal but nonetheless having exceptions. In most other sciences, one is not permitted to have arbitrary exceptions to supposedly general laws. However, in descriptive linguistics, it is clear that we very often have to make general rules to which there are then individual exceptions: in English, for instance, one can state a very general rule for the formation of the past tense of verbs or the plural of nouns – and the validity of these rules can be seen from the way in which they extend to new lexical items – yet there are still exceptions to these rules. Clearly, a universal which has no exceptions is stronger, preferable to one that does have exceptions, other things being equal, so that a priori there are arguments both in favour and against have statements of universal tendencies. It should not, however, be concluded that an absolute universal is necessarily preferable to a universal tendency, since other things are not always equal. A simple example will illustrate this. There is a nearly exceptionless universal that languages in which the relative clause precedes the head noun (RelN, i.e. the opposite from the English order in relative clause constructions) are verb-final. The known exception to this universal is Chinese, which has the orders RelN but SVO. One can therefore readily set up a universal tendency: if RelN, then SOV. However, in principle one could strengthen this by selecting any arbitrary property of Chinese and incorporating the negation of this into the *p* part of the implication; for instance, since Chinese is tonal, one could reformulate the universal as the following absolute statements: if RelN and not tonal, then SOV. Chinese is no longer an exception, since it does not fall under the first part of the implication (i.e. it is not both RelN and non-tonal). Yet whereas the original universal tendency has a certain coherence to it, in that it