Typology and Universals

Second Edition

WILLIAM CROFT

Department of Linguistics
University of Manchester

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 1990
Reprinted 1993, 1996
Second Edition 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/13 pt *System* LaTeX 2$_\varepsilon$ [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Croft, William.
Typology and universals / William Croft. – 2nd edn.
p. cm. – (Cambridge textbooks in linguistics)
Includes bibliographical references and index.
ISBN 0-521-80884-7 – ISBN 0-521-00499-3 (pbk.)
1. Typology (Linguistics) 2. Universals (Linguistics) I. Title. II. Series.
P204 .C7 2002
410′.1 – dc21 2002019247

ISBN 0 521 80884 7 hardback
ISBN 0 521 00499 3 paperback

Contents

Figures

Tables

1 Introduction

1.1 What is typology?

The term **typology** has a number of different uses, both within linguistics and without. The common definition of the term is roughly synonymous with 'taxonomy' or 'classification', a classification of the phenomenon under study into types, particularly structural types. This is the definition that is found outside of linguistics, for example in biology, a field that inspired linguistic theory in the nineteenth century.

The most unassuming linguistic definition of typology refers to a classification of structural types across languages. In this definition, a language is taken to belong to a single type, and a typology of languages is a definition of the types and an enumeration or classification of languages into those types. We will refer to this definition of typology as **typological classification**. The morphological typology of the nineteenth and early twentieth centuries is an example of this use of the term. This definition introduces the basic connotation that the term typology has in contemporary linguistics: typology has to do with **cross-linguistic comparison** of some sort. Methodological issues in cross-linguistic comparison will be discussed in §§1.3–1.6, while chapter 2 will be devoted to the notion of a linguistic type, including morphological typology, and its refinements in twentieth-century research.

A second linguistic definition of typology is the study of patterns that occur systematically across languages. We will refer to this definition of typology as **typological generalization**. The patterns found in typological generalization are language **universals**. The classic example of a typological universal is the implicational universal. An example of an implicational universal is the generalization, 'if the demonstrative follows the head noun, then the relative clause also follows the head noun.' This universal cannot be discovered or verified by observing only a single language, such as English. One has to do a general survey of languages to observe that the language type excluded by the implicational universal – namely a language in which the demonstrative follows the head noun and the relative clause precedes it – indeed does not exist.

Typological generalization is generally regarded as a subdiscipline of linguistics – not unlike, say, first language acquisition – with a particular domain of linguistic facts to examine: cross-linguistic patterns. Typology in this sense began in earnest with Joseph H. Greenberg's discovery of implicational universals of morphology and word order, first presented in 1960 (Greenberg 1966a). The primary purpose of the present volume is to discuss the kinds of cross-linguistic patterns that have been discovered and the methodological and empirical issues raised by the study of these patterns. Chapters 3–7 are devoted to discussing these patterns and the empirical and methodological issues that their discovery raises. The kinds of cross-linguistic patterns actually found represent a coherent set of language universals which are basic phenomena to be explained by any linguistic theory.

The third and final linguistic definition of typology is that typology represents an approach or theoretical framework to the study of language that contrasts with prior approaches, such as American structuralism and generative grammar. In this definition, typology is an approach to linguistic theorizing, or more precisely a methodology of linguistic analysis that gives rise to different kinds of linguistic theories than found in other approaches. Sometimes this view of typology is called the Greenbergian, as opposed to the Chomskyan, approach to linguistic theory (after their best known practitioners; see, for example, Smith 1982:256). This view of typology is closely allied to **functionalism**, the view that linguistic structure should be explained primarily in terms of linguistic function (the Chomskyan approach is contrastively titled **formalism**). For this reason, typology in this sense is often called the **(functional–)typological approach**, and will be called so here. More precisely, we may characterize this definition of typology as **functional–typological explanation**. The functional–typological approach became generally recognized in the 1970s; important figures beginning at that time include Givón, Haiman, Comrie, Hopper and Thompson. Functional–typological explanation has well-established historical antecedents, however (see Haiman 1985 and chapter 9), not least Greenberg himself.

The three linguistic definitions of typology correspond to the three stages of any empirical scientific analysis. Typological classification represents the observation of an empirical phenomenon (language) and classification of what we observe. Typological generalization – language universals – is the formation of generalizations over our observations. And the functional-typological approach constructs explanations of the generalizations over what we have observed. In this sense, typology represents an **empirical scientific** approach to the study of language.

Of course, in any empirical science the actual process of doing science does not proceed in these three discrete stages. In particular, explanations offer themselves at all stages in the scientific process. We will present typological explanations of language universals as the universals themselves are introduced in chapters 3–7. The explanatory models used by typologists include competing motivations, economy, iconicity, processing, semantic maps in conceptual space, and a rethinking of syntactic argumentation. One significant dimension of typological explanation is that explanations of many grammatical phenomena are fundamentally diachronic, not synchronic. The diachronic approach requires a fundamental rethinking of typological principles, and is discussed in chapter 8. Chapter 9 then summarizes the approach to language that typology presents.

Not surprisingly, these differing definitions of typology – typological classification, typological generalization and functional–typological explanation/approach – have led to some confusion about what typology is, or is supposed to be. For example, it is sometimes claimed that typology is ‘merely descriptive’ or ‘taxonomic’; that is to say, it does not provide a means for developing theories of language which can function as an alternative to, for example, generative linguistic theory. This represents a confusion of typological classification with typological generalization and explanation. Typological generalization represents a well-established method of analysis, and the typological approach is now a well-articulated approach to language.

The emphasis on theory and methodology in this volume should not be interpreted as minimizing the descriptive work necessary to develop typological analyses. The descriptive work which has been and, I hope, will continue to be done on the tremendous number of languages in the world is absolutely essential not just to typological theory but to all linguistic theories. Unfortunately, typological studies have often had to withhold or remove their data sections upon publication due to size limitations,[1] while many good descriptive works such as the University of Hawaii Press PALI series of Micronesian language grammars rapidly go out of print. The attitude that descriptive work is not valued (it is ‘merely’ descriptive or, disparagingly, ‘descriptivist’) must be abandoned for there to be progress in linguistic theory.

This matter becomes even more urgent because of the alarming loss of the empirical data base for linguistic theory. Hundreds of languages have become extinct in the last century. Hundreds, perhaps thousands, of others no longer survive in viable speech communities; the languages are dying and there are often serious consequences affecting grammatical structure. This situation is getting worse, not

[1] On some occasions, the data is published elsewhere. The data for Keenan and Comrie’s study on the Noun Phrase Accessibility Hierarchy (Keenan and Comrie 1977; see chapter 5) was eventually published in another journal (Keenan and Comrie 1979); the data from Maxwell’s study on linearization (Maxwell 1984) was published by a linguistics department (Maxwell 1985); and the data on Kortmann’s study of adverbial subordinators in European languages (broadly construed; Kortmann 1997) was published on diskette by LINCOM Europa.

better, and is finally achieving the attention it deserves (Dorian 1981; Krauss 1992; Crystal 2000; Nettle and Romaine 2000). The empirical problems with language research parallel the problems in biological research, in particular in evolutionary theory and ecology: the extinction of languages and the loss of the linguistic communities is like the extinction of species and the loss of their habitat (ecosystems). In both disciplines it threatens theoretical progress.

1.2 Typology, universals and generative grammar

Greenberg's approach to language universals emerged at about the same time as Chomsky's, in the late 1950s. The conception of language universals in typology and generative grammar is quite different. In this section, we will briefly describe the emergence of Greenberg's and Chomsky's ideas, and the similarities and differences that are found in the two approaches to language (for more detailed discussion, see Hawkins 1988). We will return to the relationship between typology and generative grammar in later chapters in the context of more specific theoretical issues (see §§3.5, 7.2, 9.2–9.3).

Language universals reflect the belief that there exist linguistic properties beyond the essential definitional properties of language that hold for all languages. Although this belief has considerable modern currency, it is by no means a necessary fact or universally-held opinion, and in fact the opposite view was widely held until around 1960. To a considerable degree, the difference between the generative and typological approaches to language universals can be traced to the different traditions to which Chomsky and Greenberg responded. The generative approach represents a reaction against behavioristic psychology, while the typological approach represents a reaction against anthropological relativism.

The behaviorist view of language, in particular language learning, is anti-universalist in that it posits no innate, universal internal mental abilities or schemas. In the behaviorist view, linguistic competence is acquired through learning of stimulus–response patterns. In contrast, the generative approach posits the existence of innate internal linguistic abilities and constraints that play a major role in the acquisition of language. It is these constraints that represent linguistic universals in this approach. The argument used by Chomsky (e.g. Chomsky 1976) for the existence of innate universal linguistic competence refers to the 'poverty of the stimulus'. It is argued that the child has an extremely limited input stimulus, that is, the utterances that it is exposed to from the mother and other caregivers. This stimulus is incapable of permitting the child to construct the grammar of the adult's language in a classic behaviorist model; therefore, the child must bring innate universals of grammatical competence to bear on language acquisition. Hence, the primary focus on universals in the generative tradition has been on their innate character.

The anthropological relativist view of language is that the languages of the world can vary arbitrarily: ‘languages could differ from each other without limit and in unpredictable ways’, in Martin Joos’ famous passage (Joos 1957:96). This view of language was particularly strong among anthropological linguists studying North American Indian languages, which indeed differ radically in many ways from so-called Standard Average European languages. However, the comparison of one ‘exotic’ language or a limited number of languages to English only indicates diversity, not the range of variation, let alone limits thereto. Greenberg and others discovered that a more systematic sampling of a substantial number of languages reveals not only the range of variation but constraints on that variation. Those constraints demonstrate that languages do not vary infinitely, and the constraints represent linguistic universals. Hence, the primary focus on universals in the typological tradition has been on their cross-linguistic validity, and on universals that restrict possible language variation (see §3.1).

The innate universals posited by generative grammar are intended to explain linguistic structure. The poverty of the stimulus argument is essentially a deductive argument from first principles (although it does make assumptions about the nature of the empirical input, and what counts as relevant input). The poverty of the stimulus argument is one aspect of Chomsky’s more generally **rationalist** approach to language. The universals posited by typology are intended to represent inductive generalizations across languages, in keeping with typology’s **empiricist** approach to language. Typological universals call for explanation in terms of more general cognitive, social-interactional, processing, perceptual or other abilities. These abilities may also be innate, but they extend beyond language per se. The generative grammarian argues that the discovery of innate principles that the child brings to bear in learning a single language can be extrapolated to language in general (Chomsky 1981). The typologist argues that a grammatical analysis based on one language or a small number of languages will not suffice to reveal linguistic universals; only a systematic empirical survey can do so.

These differences in approach have led to claims that the Greenbergian approach and the Chomskyan approach to language universals and linguistic explanation are diametrically opposed to each other. In fact, there are significant similarities between the generative and (functional–)typological approaches. Both approaches begin with the analysis of language structure. Both approaches consider the central question of linguistics to be ‘What is a possible human language?’ (though see §§3.1, 8.1). Both approaches are universalist, in contrast to their predecessors. There is broad agreement that there do exist a substantial number of universals that hold of all languages (assuming attested exceptions can be accounted for by other principled factors). For both approaches, the construction of linguistic generalizations involves abstraction over the data, though the Greenbergian abstracts patterns across languages and the Chomskyan abstracts patterns within languages (see §9.2). Likewise, explanations for linguistic universals rest on universal human abilities, which may or may not be language specific, and which probably have a significant innate component, though perhaps are not entirely innate. In fact, for both generative and typological approaches, the foundations of linguistic explanation are ultimately biological, although for the Chomskyan the biological basis is found in genetics (innate linguistic knowledge) and for the Greenbergian the biological basis is indirect, and is to be found in evolutionary theory (see §9.3; Croft 2000).

Nevertheless, there are two salient distinctive characteristics of the Greenbergian approach: the central role of cross-linguistic comparison, and the close relationship between linguistic form and language function. These two characteristics are discussed in the following two sections.

1.3 Cross-linguistic comparison

The first question that may be asked of typology is, what is the role of cross-linguistic comparison – the fundamental characteristic of typology – in linguistic analysis? Cross-linguistic comparison places the explanation of linguistic phenomena in a single language in a new and different perspective. For example, the distribution of the definite and indefinite articles in English is fairly complex:

(1a) He broke **a vase**.
(1b) He broke **the vase**.
(1c) The concert will be on **Saturday**.
(1d) He went to **the bank**.
(1e) I drank **wine**.
(1f) The French love **glory**.
(1g) He showed **extreme care**.
(1h) I love **artichokes** and asparagus.
(1i) Birds have **wings**.
(1j) His brother became **a soldier**.
(1k) **Dogs** were playing in the yard.

The eleven sentences given above characterize eleven types of uses of the articles (or their absence) in English, given as follows:

(a) specific (referential) indefinite (see §5.2);
(b) specific and definite;
(c) proper name;

(d) specific manifestation of an institution/place;
(e) partitive of a mass noun;
(f) generic mass noun;
(g) specific manifestation of an abstract quality (mass noun);
(h) generic of a count noun;
(i) generic of an indefinite number of a count noun;
(j) predicate nominal;
(k) specific but indefinite number of a count noun.

It might be possible to develop a set of generalizations – an **analysis** – that predicts exactly the distribution of the two articles (including their absence) in English. Such an account may be syntactic, semantic or pragmatic, or a combination of all three. Whatever is the case, it will have to be a fairly complex and subtle analysis, especially since the eleven different construction types given here do not exhaust the possibilities.

At this point, the typologist will ask: what is the significance of these generalizations posited in English for the class of human languages as a whole? Examining even a relatively closely related language, French, produces difficulties for those generalizations. In the exact same contexts, illustrated here by translation equivalents of the English sentences, the distribution of definite and indefinite articles *le/la/les* and *un/une* respectively (and their absence) is quite different:

(2a) Il a cassé **un vase**.
(2b) Il a cassé **le vase**.
(2c) Le concert sera **samedi**.
(2d) Il est allé à **la banque**.
(2e) J'ai bu **du vin**. (du = de + le)
(2f) Les Français aiment **la gloire**.
(2g) Il montra **un soin** extrême.
(2h) J'aime **les artichauts** et les asperges.
(2i) Les oiseaux ont **des ailes**. (des = de + les)
(2j) Son frère est devenu **soldat**.
(2k) **Des chiens** jouaient dans le jardin.

It is quite likely that the analysis of the distribution of the English articles would have to be drastically altered if not abandoned and a new one developed for the distribution of the French ones. In French, we find a more widespread use of both the French definite and indefinite articles, the appearance of the partitive marker *de* plus the definite article, and the absence of the French indefinite article in the predicate nominal construction.

One cannot be certain how much we would have to start all over again, of course, since to the best of my knowledge no complete analysis has been worked out. However, a generalization for a subset of three of the eleven contexts has been proposed, for the generic count nouns in (h) and (i) and the indefinite number of count-noun usage in (k). Carlson (1977) proposes a unified analysis of the bare plural construction used in both situation types, in which both are of the same semantic type and the differing interpretations are attributed to the semantic type of the predicate. But when we turn to French, we see that in fact two different types of constructions are found – compare 2h and 2i,k – and so this generalization does not clearly apply to the grammatical facts of French. One may try to attribute the difference to the French partitive marker *de*. But if we turn to still other languages such as Rumanian (Farkas 1981:40–45), which distinguish the two uses solely by the presence vs. absence of the article, then we will not be able to invoke such an alternative.

The fact that analyses of linguistic phenomena 'one language at a time' cannot be carried over from one language to the next is somewhat disturbing for the search for language universals. Intricate interactions of internal structural generalizations are proposed by linguists to 'predict' grammatical patterns that do not apply even to neighboring languages. This is true not only in structuralist–generative analyses. Functionalist analyses, which invoke external (semantic or pragmatic) generalizations to account for the distribution of phenomena like the articles of English, often have the same problems:

> Volumes of so-called functionalism are filled with ingenious appeals to perception, cognition or other system-external functional domains, which are used to 'explain' why the language in question simply has to have a grammatical particularity that it does – when a moment's further reflection would show that another well-known language, or even just the next dialect down the road, has a grammatical structure diametrically opposed in the relevant parameter. (DuBois 1985:353)

The question here is, to what level of generalization should an analysis of language-specific facts be developed before taking into consideration cross-linguistic patterns? The typologist essentially takes the position that cross-linguistic patterns should be taken into consideration at virtually every level of generalization about human languages (see §9.3).

A cross-linguistic comparative approach – that is the construction of typological generalizations – allows us to make progress on universal characteristics of the distribution of articles, for example, and in turn causes us to reassess an analysis formulated without reference to the facts in other languages.

There are certain generalizations that cut *across* the two languages that are very likely to be characteristic of language in general. For instance, the first three uses, (a)–(c), are identical in English and French, and it is only in the following seven that there is substantial variation between the two languages. With the exception of the (k) use, all of the variable uses across the two languages concern generic and mass-noun contexts of various sorts. This suggests that there may be some degree of uniformity across languages in specific NP contexts that does not exist in generic and mass NP contexts. (In fact, there is also variation in specific NP contexts, but of a more constrained type; see §8.2.)

There are two important points implicit in this proposed generalization over the English and French facts which summarize the argument for cross-linguistic comparison. The first is that this generalization could not be formulated without looking at more than one language. (Examining still more languages would, of course, further refine this generalization.) That is what makes this analysis of the grammatical phenomenon typological.

The second point pertains to the description and analysis of the grammar of a particular language, given the sorts of cross-linguistic generalizations that exist. Awareness of cross-linguistic variation allows the linguist describing a particular language to provide a more fine-grained description of the phenomenon in question. For example, being aware of the differences between English and French in generic and mass-noun contexts implies that a grammatical description should explicitly indicate how a language with articles expresses or **encodes** those different semantic types of NPs.

A fine-grained description of the linguistic facts of a language is sufficient for descriptive completeness. Of course, one always wants to seek generalizations in the data. Moreover, one would like the generalizations to correspond to some empirically real phenomenon, such as a speaker's knowledge of her (or his) language. If the generalizations are intended to represent a speaker's knowledge of her language, then such an analysis must integrate cross-linguistic comparison, according to the typological approach. For example, the generalizations about the distribution of the articles in both English and French ought to characterize the distribution in specific NP contexts in each language as typical or even universal (if that turns out to be the case), and the distribution in generic and mass NP contexts as arbitrary and language specific, or perhaps subject to other conditions that would be revealed by further cross-linguistic comparison. In this view, the analysis of the articles in French or English would be incomplete – and therefore an inadequate explanation of the phenomenon – if its relationship to cross-linguistic generalizations about articles is not taken into account. The generalizations revealed by examining more than one language at a time are the only ones which can be said to hold of languages in general. A speaker's knowledge of her language involves both universal and language-particular properties.

Until relatively recently, typology has not directed its attention to the relationship between language universals and the generalizations posited in particular language grammars (Croft 1999; §9.1). However, it is not the case that language universals exist independently apart from the linguistic knowledge of language speakers. More recent typological research has begun to address this question, and has developed models of representing language-particular facts and language universals (see in particular §5.3).

Another illustration of the need for a cross-linguistic approach in formulating linguistic generalizations, and the difference between a cross-linguistic approach and a 'one language at a time' approach, is found in syntactic argumentation. Syntactic arguments are constructed by means of the **distributional method**: one examines the occurrence or **distribution** of a grammatical category in a series of different constructions, and the existence of the category is justified if the distribution pattern is the same across the constructions. For example, in arguing for the category subject in English, the distribution of the immediately preverbal NP is the same in the constructions illustrated in 3–7, and justifies its categorization as the subject:

(3) **He**/*him congratulated him.
(Nominative case of the pronoun *he* as opposed to *him*.)
(4) Teresa **likes**/*like horses.
(Agreement of the verb with *Teresa*.)
(5) Jack$_i$ wants **Ø$_i$** to leave.
(The person understood to be leaving is Jack; Jack controls the unexpressed argument of the infinitive following *want*.)
(6) **Ø$_i$** Take out the garbage.
(The unexpressed argument in the imperative construction.)
(7a) **John$_i$** found a ring and **Ø$_i$** took it home with him.
(7b) *John found a ring$_i$ and Ø$_i$ was gold.
(The unexpressed shared argument in a conjoined sentence.)

In terms of standard syntactic argumentation, 3–7 give five independent pieces of evidence for identifying the immediately preverbal NP as the subject of the clause. Another way of putting it is that positing the existence of a category subject in English is constructing a generalization over the distributional facts presented in 3–7.

A typological analysis, on the other hand, would not present the preceding facts as arguments for a category subject. The facts in 3–7 are a generalization formed by examining just one language. For the typologist, the significant questions all refer to the status of this correlation cross-linguistically. Again, we must ask: What elements of this correlation are accidental, a peculiarity of English? What elements of this correlation are universal? What correlations systematically vary across languages, and why? (An answer to these questions for the constructions above, other than 6, is presented in §7.1.)

The other side of the coin in invoking cross-linguistic comparison is that by examining a number of diverse languages, one will find striking, fascinating and sometimes mysterious connections between certain linguistic structures that one would not have imagined if one's attention were restricted to one language or a few typologically similar languages. This may take the form of a peculiar fact of one language which turns out to be widespread, or of a connection between two linguistic phenomena that is widespread but not manifested in one's own language.

An example of the former is the apparently arbitrary irregularity of the objective forms of the English pronouns (*me, us, him, her, them* vs. invariant *it*). This irregularity is actually a manifestation of an extremely widespread pattern of relationship between case marking and animacy, namely that direct objects that refer to more highly animate beings are more likely to have distinct object case forms (see §6.3.1). The lack of an objective form for *you* is an apparent exception which also may be due to a general pattern, namely the typological markedness of plural forms (see chapter 4; *you* was originally the second person plural form).

Another example is the variety of uses of the preposition *with* illustrated in the following sentences:

(8) I went to New York **with** John. (comitative)
(9) He opened the door **with** a crowbar. (instrument)
(10) He swims **with** ease. (manner)

Intuitively, there seems to be little if any semantic connection between these three distinct uses of the same preposition, but a typological study of the distribution of adposition/case uses reveals that the subsumption of these and certain other uses under the same adposition or case marker is actually quite common (Croft 1991a:184–92; Stolz 1996). Consider for example Hausa *dà* and Classical Mongolian *-iyer* ~ *-iyar* in the following examples:

Hausa (Abraham 1959:22; Kraft and Kirk-Greene 1973:85)

(11)	nā	hàrbē	shī	**dà**	bindingà
	1SG.COMP	shoot	3SG	**with**	gun

'I shot him with a gun.'

(12)	mun	ci	àbinci	tā̀re	**dà**	shī
	1PL.COMP	eat	food	together	**with**	3.SG

'We ate food with him.'

(13)	yā	gudù	**dà**	saurī
	3SG.COMP	run	**with**	speed

'He ran fast ["with speed"].'

Classical Mongolian (Poppe 1974:153–54)

(14)	küol	**-iyer**	giški-
	foot	**-with**	tread.on-

'to tread on with the foot'

(15) manu morin tegün -ü morin **-iyar** belčimüi
1PL.GEN horse that.3SG -GEN horse **-with** grazes
'Our horse grazes with his horse.'

(16) türgen **-iyer** yabumui
speed **-with** goes
'He goes fast.'

Investigation of this cross-linguistic phenomenon suggests that the connection between these three uses and certain other uses can be defined in terms of causal relations between participants and properties of an event (Croft 1991a: chapters 4–5).

An example of two apparently unrelated constructions in English hiding a cross-linguistically evident connection is found with conditionals and topics (Haiman 1978a). The antecedent (protasis) of a conditional sentence is marked with *if* while a sentence topic is marked in English with *as for* or *about*:

(17) **If** you eat that, you will get sick.
(18) **As for** Randy, he's staying here.

Haiman discovered that in fact conditional protases and topics are encoded identically in many languages. For example, in the Papuan language Hua a suffix -*mo* is attached to both (potential) sentence topics and conditional protases; Turkish -*sA*, the conditional marker, can also mark the contrastive topic, and the Tagalog word for 'if', *kung*, can mark contrastive topics in conjunction with the preposition *tungkol* 'about' (Haiman 1978a:566, 577). This somewhat mysterious connection between conditional protases and sentence topics found in several languages – though not in English – led Haiman to the discovery that the two are actually quite closely related in semantic and pragmatic terms.

Finally, cross-linguistic examination may also suggest that a phenomenon found in well-known languages is actually extremely unusual if not unique in the world's languages and thus may be a rather peripheral linguistic phenomenon from a cross-linguistic perspective. For example, the use of the indefinite article in predicate nominals illustrated in 1j is quite uncommon, and the phenomenon of stranding prepositions as in *the book that I told you about* is extremely rare among the languages of the world. The obligatory presence of unstressed pronouns in subject position in English is also quite uncommon cross-linguistically (see §3.5). This is not to say that such phenomena do not need to be accounted for. It is just that they are perhaps not of as great importance to the study of language universals as the more widespread or universal phenomena, such as the extremely widespread use of the bare noun for predicate nominals and the subsumption of certain case roles under the same adposition or case affix.

1.4 The problem of cross-linguistic comparability

The characteristic feature of linguistic typology is cross-linguistic comparison. The fundamental prerequisite for cross-linguistic comparison is cross-linguistic comparability, that is the ability to identify the same grammatical phenomenon across languages. One cannot make generalizations about subjects across languages without some confidence that one has correctly identified the category of subject in each language and compared subjects across languages. This is in fact a fundamental issue in all linguistic theory. Nevertheless, this problem has commanded remarkably little attention relative to its importance for linguistic theorizing.

Greenberg's original paper on word order offers the basic answer to the problem of cross-linguistic comparability:

> It is here assumed, among other things, that all languages have subject–predicate constructions, differentiated word classes, and genitive constructions, to mention but a few. I fully realize that in identifying such phenomena in languages of differing structure, one is basically employing semantic criteria. There are very probably formal similarities which permit us to equate such phenomena in different languages . . . The adequacy of a cross-linguistic definition of "noun" would, in any case, be tested by reference to its results from the viewpoint of the semantic phenomena it was designed to explicate. If, for example, a formal definition of "noun" resulted in equating a class containing such glosses as "boy," nose," and "house" in one language with a class containing such items as "eat," "drink," and "give" in a second language, such a definition would forthwith be rejected and that on semantic grounds. (Greenberg 1966a:74)

These remarks summarize the essential problem and a general solution. The essential problem is that languages vary in their structure to a great extent: indeed, that is what typology (and, more generally, linguistics) aims to study and explain. But the variation in structure makes it impossible to use structural criteria, or only structural criteria, to identify grammatical categories across languages. If we did use structural criteria, we would be prejudging the result of our supposedly empirical analysis, by excluding a priori structural types that do not fit our criteria. Hence, the ultimate solution is a semantic one.

Greenberg's remarks are echoed by Keenan and Comrie in their analysis of relative clauses in their pioneering work on noun-phrase accessibility:

> We are attempting to determine the universal properties of relative clauses (RCs) by comparing their syntactic form in a large number of languages. To do this it is necessary to have a largely syntax-free way of identifying RCs in an arbitrary language. Our solution to this problem is to use an essentially semantically based definition of RCs. (Keenan and Comrie 1977:63; see also Downing 1978:377–80, and more generally Stassen 1985:14)

In the case of relative clauses, the variation of morphosyntactic expression is such that a number of languages use morphological rather than syntactic means for forming what we would intuitively, that is semantically, identify as relative clauses (Comrie 1989:143).

The term 'semantic' as usually understood is in fact too narrow a description. Various pragmatic features, such as discourse structure (for comparing everything from forms of greeting, to discourse-defined connectives such as *anyway*, to the information structure of clauses) and conversational context (as in expressions of politeness and interlocutor status) also play a role in determining the cross-linguistic identification of the morphosyntactic phenomena that linguists are concerned with. Semantics is also irrelevant for phonological comparison. For cross-linguistic comparison of sound structure, one must base the analysis on phonetic realization (see below). These parameters are all essentially **external**, that is outside the syntactic, morphological and phonological structure of the language itself. Hence, the solution to the problem of cross-linguistic comparability is to use external definitions of grammatical categories (but see below).

Recognition of the problem of cross-linguistic comparability and its solution has led to the formulation of a standard research strategy for typological research:

(i) Determine the particular semantic(-pragmatic) structure or situation type that one is interested in studying.

(ii) Examine the morphosyntactic construction(s) or **strategies** used to **encode** that situation type.

(iii) Search for dependencies between the construction(s) used for that situation and other linguistic factors: other structural features, other external functions expressed by the construction in question, or both.

This solution to the problem of cross-linguistic comparability implies a close relationship between form and external function. Typological classification – the descriptive prerequisite to typological generalization and explanation – requires a cross-linguistic analysis of the relationship between linguistic form and function. Since this is a controversial point in contemporary linguistic theory, it is worth examining the problem more closely.

Many grammatical categories are identified cross-linguistically by semantic means without significant objections. For instance, if one is trying to find out if a set of verbal suffixes represents tense or aspect, one examines their meaning and use, not any formal properties. In these categories, difficulties in cross-linguistic comparability arise chiefly when a single form combines multiple functions (as often happens). The main problematic categories for cross-linguistic identification are the fundamental grammatical categories: noun, verb and adjective, subject and object, head and modifier, argument and adjunct, main clause and subordinate clause, etc. (Croft 2001). Needless to say, these categories are central to linguistic theory. On the one hand, these categories do not have an obvious functional (semantic and/or pragmatic) definition. On the other hand, these grammatical categories and the categories defined by them do vary considerably in their structural expression across languages, once we have identified the categories cross-linguistically by semantic/pragmatic means.

The problem of cross-linguistic identification should not be overstated. In most cases it is not difficult to identify the basic grammatical categories on an intuitive basis. To a great extent this is accomplished by examining the translation of a sentence and its parts, which is of course the semantic/pragmatic method. On the other hand, the weaknesses of an intuitive cross-linguistic identification of categories become apparent when one focuses on an example which is not so intuitively clear after all (for example, is the English gerund form in *Walking the dog is a chore* a noun or a verb?).

To give an idea of how unavoidable considerations of external function are, we will briefly discuss some of the problems involved with a cross-linguistic identification of subject. First, across languages, the grammatical relation of subject is expressed structurally in several different ways: by case-marking (including adpositions), by indexation (agreement), by word order, or by a combination of both of these. Yet, how does one know this in the first place? Only by using a cross-linguistic definition involving external function, including some notion of agent of an action and topic of the sentence, to determine what is a subject in each language.

Now one must have a cross-linguistic means to identify case/adposition, indexation and word order. Word order appears to be the easiest, since it is clearly based on a physical property of the utterance, the sequence of units, which can be directly observed. However, the correct word order analysis requires that the grammatical category of each unit be identified. For example, the assertion that Yoruba subjects can be identified by their position before the verb requires the identification of verbs in Yoruba, not to mention noun phrases or at least nouns (and not to mention a cross-linguistic means of individuating syntactic units, a problem that we will not deal with here).

A cross-linguistic definition of case/adposition and indexation on a structural basis is difficult as well. Case/adposition markers can be attached to the NP argument, or be independent particles, or even be attached to the verb in some cases, so syntactic position and dependency cannot be suitable criteria for a cross-linguistic definition.

Attachment to subject: Russian

(19) pis'm **-o** lež -it na stol -e
letter **-NOM**.SG.N lie -3SG.PRS on table -LOC.SG.M.
'The letter is lying on the table.'

Independent particle: Rumanian (Nandris 1945:145)

(20) pune cartea **pe** masă!
put:IMP book:DEF **on** table
'Put the book on the table!'

Attachment to verb: Mokilese (Harrison 1976:164)

(21) Ngoah insingeh **-ki** kijinlikkoau -o nah pehnn -o
I write.TRNS **-with** letter -DET his pen -DET
'I wrote that letter with his pen.'

Indexation markers (agreement; see §2.1.3) are syntactically at least as variable: they can be affixes to the verb, independent particles, or attached to other constituents of the sentence, including the noun phrase denoting the same referent as the index:

Attachment to verb: Hungarian (Whitney 1944:15)

(22) áll **-unk**
stand **-1PL**.INDF
'We are standing.'

Independent particle: Woleaian (Sohn 1975:93)

(23) Sar kelaal **re** sa tangiteng
child those **3SG** ASP cry.RDP
'Those children over there cried and cried.'

Attachment to other constituents: Ute (first constituent; Givón 1980a:311)

(24) kavzá -yi **-amu̧** -'ura mag̑á -x̂a -páa-ni
horse -OBJ **-3PL** -be feed -PL -FUT
'They are going to feed the horse.'

Attachment to any constituent, including noun phrase: Bartangi (Payne 1980:163, 165; compare Santali, cited in Sadock 1991:146)

(25) a̅z **-um** ta̅ -r kitob vuj
I **-1SG** you -to book bring.PRF
'I have brought you a book.'

Thus, morphosyntactic dependence – e.g. case marking on subjects, agreement on verbs – will not provide an unproblematic cross-linguistic definition, at least not by itself. A more suitable definition would be that a case marker/adposition is **relational**, that is, a morpheme that denotes the semantic relation that holds between the noun phrase and the verb, while agreement is **indexical**, that is, a morpheme that denotes the argument itself (Croft 1988; §2.1). This definition is essentially a semantic one.

If we assume a cross-linguistic definition for case marking and indexation that fits our intuitions, then we encounter a larger problem. Our intuitive notion of subject is based on English subjects (or Standard Average European subjects, to use Whorf's [1956:138] term), specifically, on the semantic relation between the event denoted by the verb and the participant denoted by the English subject. Examining more 'exotic' languages, we find that what we have identified as the subject by the use of a particular case-marking or indexation form does not correspond to English subjects, or the English subject does not conform to the other language's subject. For example, in Chechen-Ingush (Nichols 1984:186), the English translation 'subjects' of the following three examples display quite different case-marking and indexation patterns.

(26) bier **-Ø** **d-** ielxa
child **-NOM** **CL-** cries (CL indexes 'child')
'The child is crying.'

(27) a:z yz kiniška -Ø **d-** ieš
1SG.**ERG** this book -NOM **CL-** read (CL indexes 'book')
'I'm reading this book.'

(28) suona yz kiniška -Ø **d-** iez
me.**DAT** this book -NOM **CL-** like (CL indexes 'book')
'I like this book.'

If we identify the subject with the nominative noun phrase that the verb indexes, then 'this book' in the second and third sentences is the subject. If we treat the ergative and/or dative noun phrase as subject, then the first sentence appears not to have a subject. Whatever solution is taken to this problem must refer to the actual semantic relations that hold between the subject and the verb (§5.4). Thus, a cross-linguistically valid definition of subject referring to external properties is unavoidable.

It is possible to develop cross-linguistic definitions of grammatical categories that are partially structural in nature. Many grammatical constructions are defined in terms of the basic grammatical categories whose difficulties in cross-linguistic identification we have discussed: subject, noun, verb, etc. If these basic categories can be identified across languages by external definitions, one may develop **derived structural** definitions for the construction in question. For example, the passive construction can be defined as one in which the subject of the passive verb is the object of the counterpart active verb. This is a structural definition of the passive that can be used for cross-linguistic identification, once one has already identified subject, verb, object and the active construction on external grounds.

The choice of a purely external vs. a derived structural definition of a construction depends on the purposes of the typological study. For example, we may compare external and derived structural definitions for the subjunctive. An external definition is that the situation denoted by the subjunctive clause is nonfactual or irrealis modality. A derived structural definition is that a subjunctive clause is a clause which (1) expresses the subject and the object of the clause in the same way as an ordinary declarative main clause does, but (2) whose verb inflections differ from those of the verb in an ordinary declarative main clause. Condition (1) is intended to distinguish the subjunctive from various types of nonfinite clauses; and condition (2) is intended to distinguish the subjunctive from the indicative. The external definition would be more useful in a typological study of modality (e.g. Palmer 1986; Bybee, Perkins and Pagliuca 1994); but the derived structural definition has proved more useful in studies of complex sentence structure (e.g. Stassen 1985; Koptjevskaja-Tamm 1993; Croft 2001: chapter 9; Cristofaro 2003).

Not all externally based definitions are created equal. For example, in seeking a cross-linguistically valid definition of subject, one would not use translation equivalents of expressions such as ‘The lightning struck the tree’ or ‘I like bananas’; one would more likely use expressions like ‘I broke the stick’ or ‘He killed the goat.’ A priori, there is no reason to select the latter two as better for defining subject than the former two. In either case, one determines the grammatical relation of the relevant predicate–argument relation and identifies it as the subject. Nevertheless, our pretheoretic intuitions about grammatical categories strongly suggest that some external definitions are better cross-linguistic criteria than others, and detailed analysis of the relevant linguistic phenomena generally bears out those intuitions. Hence, we use physical actions with animate agents for defining subject, the relationship of ownership for defining (alienable) possession, and so on.

Of course, these choices are based on pretheoretic intuitions and may turn out to be incorrect. For example, many linguists, e.g. Faltz (1978), use the recipient of the verb “give” as the defining environment for the dative, but others have argued that ‘in many languages . . . “give” is syntactically a very atypical ditransitive verb . . . selection of “give” always requires cross-checking with a variety of other verbs of similar valency’ (Borg and Comrie 1984:123). Reliance on single exemplars can lead to building a typological generalization on too narrow an empirical base. What matters are the cross-linguistic facts. The best external definitions are those that yield categories with more consistent coding across languages and more consistent grammatical behavior (distribution patterns). In fact, a cross-linguistic study must be somewhat broad in semantic range in order to discover the best cross-linguistic definitions.

In phonology, the problem of cross-linguistic comparison also arises. The cross-linguistic identification of English /p/ with Russian /p/ is based primarily on their articulatory-acoustic similarity, that is, their external, phonetic, values. Also, to argue that a category [p] participates in a typological pattern involving a hierarchy of stops including [t] and [k] (see §5.5) presumably means that the articulatory gestures and/or acoustic features involved in [p] are related to those involved in [t] in such a way as to manifest the linguistic behavior which led us to postulate the hierarchy in the first place. It is difficult to see how one could use any other criterion, because there may be no obvious way to identify anything in the alternative phonemic system with English /p/, because of differences between the other phonemic system and the English system.

For example, if the language is Hindi, which distinguishes between aspirated /p^h/ and unaspirated /p/, the whole phonemic system is different, and so it would be impossible to identify English /p/ with Hindi /p/ on the basis of the phonemic system. The problem is, it is extremely difficult to gauge which Hindi phoneme the English /p/ should be identified with phonetically. Most allophones of English /p/ are quite aspirated like Hindi /p^h/; but those allophones do not contrast with phoneme /p/, unlike Hindi /p^h/. Most phonological typological studies have involved the analysis of phoneme inventories, making generalizations based, for example, on five-vowel systems vs. seven-vowel systems. However, phonetically, not all seven-vowel systems are alike; the individual vowels differ acoustically, and this is generally true for all phonological segments (Ladefoged and Maddieson 1996). In a typological approach to phonology, one must also do cross-linguistic comparison on the basis of the relationship between the linguistic system and its external (phonetic) manifestation.

1.5 Language sampling for cross-linguistic research

There are approximately six thousand languages in the world today. The majority of them are not documented at all, or have only minimal documentation (e.g. word lists). Of the rest, documentation varies substantially in quality. Even so, there are hundreds of languages for which good documentation is available. If one is examining a phenomenon which is exhibited in only a relatively limited number of languages – such as implosives or numeral classifiers – then one can examine virtually all attested examples. For example, in his study of glottalic consonants, Greenberg (1970) used a sample of 150 languages which was virtually exhaustive at the time for implosives (though not for ejectives). But in most cases, the available documentation is far greater than can be handled in most realistic typological studies. Hence, typologists must use a subset of these languages in studying cross-linguistic variation, that is, a **sample**. But by using a subset of the world's languages, two methodological problems arise.

The first problem is that the sample may not capture all linguistic diversity. Consider for example the English passive voice construction:

(29) The boy **was** tak**en** to school (**by** his parents).

The English passive differs from the English active by the presence of two words, the auxiliary *be* and the preposition *by*, as well as the verbal inflection (the passive participle suffix). Other European languages have structurally identical passives, so one might venture the hypothesis that passives always involve the presence of an auxiliary and/or a preposition governing the agent phrase. In fact, this is not so. Lummi, like many languages, expresses the passive without an auxiliary, simply with a verbal inflection (Jelinek and Demers 1983:168):

(30)	x̣či	-t	-ŋ	-sxʷ	ə	cə	swəyʔəʔ
	know	-TR	**-PASS**	-2	**by**	the	man

‘You are known by the man.’

Bambara represents a much rarer case, a ‘passive’ without overt marking of the verb form as passive (Chris Culy, pers. comm.):

(31)	o	fo	-ra	dugutigi	**fè**
	3SG	greet	-CMPL.INTR	chief	**with**

‘S/he was greeted by the chief.’

The second problem is that what we think of as a theoretically significant relationship between two grammatical properties may be an accident. For example, it has been suggested that there is a biconditional relationship between the absence of obligatory independent subject pronouns and indexation (the ‘Taraldsen generalization’ [Taraldsen 1980]; see, for example, Huang 1984:534). The basis for this hypothesis is the fact that English has obligatory subject pronouns but very little indexation, whereas many European languages (so-called pro-drop or null subject languages) do not have obligatory subject pronouns but have rich indexation systems. The difference can be illustrated by comparing the English examples in 32 to the Spanish examples in 33:

(32a) **I** ate the bread.
(32b) *Ate the bread.

(33a) **Yo** comé **(Tú** comiste, **Él/Ella** comió, *etc.*) el pan.
‘I ate (you ate, he/she ate, *etc.* the bread.’
(33b) Comé (comiste, comió, *etc.*) el pan.

Nevertheless, there exist many languages – in fact more languages than the English type according to one typological survey (Gilligan 1987:131–32) – that do