

Typology and Universals

William Croft, University of New Mexico

1. Introduction: the typological and generative approaches to language universals

Typology represents an approach to the study of linguistic structure that differs in certain important respects from the generative and the functionalist approaches (Wasow, this volume; Van Valin, this volume), although it is closer in spirit to the latter. The most important difference between typology and these other approaches to linguistic structure is that the typological approach is fundamentally crosslinguistic in nature. A formalist linguist can, and most formalist linguists do, analyze a single language in the search for universals of language structure. There are formalist analyses of many different individual languages, but relatively few crosslinguistic formalist studies (a notable exception is the work of Mark Baker, e.g. Baker 2003; but see Croft 2008). A functionalist linguist can, and often does, analyze a single language in the search for universals of the relationship of language structure to language function. Some functionalist theories are more extensively supported by crosslinguistic data, notably Functional Grammar (Dik 1997) and Role and Reference Grammar (Van Valin and LaPolla 1997). These theories recognize the variation in grammatical structures across languages, but their frameworks generally focus on what is believed to be common to all languages (see §5).

Typology, on the other hand, is fundamentally comparative. A genuinely typological analysis of a grammatical construction, or a phonological pattern, or other aspects of language, examines the variation across a large number of languages. In this respect, typology resembles comparative historical linguistics. The goals of typology and comparative historical linguistics are very different, however, although the results of each are essential to the other (see §2). Comparative historical linguistics seeks genetic (family tree) relationships among languages, in order to discover the history of the languages and their speech communities (Joseph, this volume). Typology examines a broad sample of languages in order to discover universals of language structure and propose explanations of those language universals.

For this reason, typology is linked to language *universals*. For some linguists, typology simply means the description of variation, that is, how languages differ in their structure. For example, a simple descriptive typology of the word order of numeral and noun would divide the languages of the world into three broad *types*:

- Those in which numerals normally precede the noun they modify, as in English (Indo-European) *two women*;
- Those in which numerals normally follow the noun they modify, as in Ma'di (Nilo-Saharan, Uganda-Sudan) *àg̃ sū* 'men four';
- Those in which numerals may either precede or follow the noun they modify, as in Wardaman (Australian, Australia) *guyaminyi mulurruwuyayi* 'two old.women' or *marluga lege* 'old.man one'

However, typology in the linguistic sense is more than a classification of how languages differ in their structure. A descriptive typology leads to generalizations that constrain how much languages can vary; those generalizations are language universals. Several examples of language universals will be given in this chapter.

The search for language universals is shared by typology and generative grammar. However, the language universals derived from typological research are quite different from those derived in generative grammar, although the generative and typological approaches arose at around the same time. Although the belief in language universals 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 in American linguistics 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 their respective founders, Noam Chomsky and Joseph 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 language 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 a famous quotation from the linguist Martin Joos (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 language universals. Hence the primary focus on universals in the typological tradition has been on their cross-linguistic validity, and on universals that restrict language variation (see §5).

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. 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 language universals; only a systematic empirical survey can do so.

The typological approach to discovering language universals, like any empirical scientific approach, is basically *inductive*. In order to discover what language in general is like, or the universal character of language, one should look at a large number of languages. For example, consider the two English examples in (1)-(2):

- (1) They talked about the war.
- (2) What did they talk about?

The question in (2) differs from the statement in (1) in three ways: the interrogative pronoun corresponding to *the war* in (1) is found at the beginning of the question; the preposition *about* nevertheless remains "stranded" in its position after the verb; and the auxiliary verb *did* occurs, positioned before the subject. Looking only at English, one cannot tell what the significance of these three differences are. In fact, the initial position of the interrogative pronoun is extremely widespread across the world's languages, though by no means universal. In contrast, the stranding of the preposition and the insertion of an auxiliary that is absent in the corresponding statement are extremely rare and idiosyncratic grammatical traits of English, and not causally connected to each other or to the initial position of the interrogative pronoun. It is only by examining a broad sample of languages that the significance of different properties of linguistic structure can be assessed.

Likewise, one can only judge whether a causal connection exists between two grammatical properties by examining a large and widely distributed sample of languages. For example, languages differ in the constructions used for nonverbal predication. English requires a *copula* (in boldface in (3)) for nonverbal predication at all times, whereas Russian lacks a copula in the present tense:

- (3) She **is** a doctor.
- (4) Ona vrač.
 She doctor

A common explanation for the absence of a copula, or *zero* copula, in a variety of theories is that the reason for the occurrence of a copula is a "need" to place verbal inflections that cannot occur on nonverbal predicates such as the nouns in (3)-(4) (see Stassen 1997:66 for the history of this proposal, and also Baker 2003:40). Zero copulas

therefore are claimed to occur when verbs lack inflection or when the inflection is zero; nonzero inflections require a “dummy” copula to carry them. Thus the zero copula occurs in Russian present tense because that is the “unmarked” inflectional category.

This hypothesis appears to be a plausible one, assuming that the purpose of the copula is to carry verbal inflections for nonverbal predicates. It appears to be valid for English, Russian, and a number of other languages. However, Stassen demonstrates that in fact it is not crosslinguistically valid in general (Stassen 1997:65-76). There are many languages with overt inflections but a zero copula, such as Sinhalese (Indo-European, Sri Lanka; see the Appendix for abbreviations of grammatical elements in the examples):

- (5) mahattea e-nəw-a
gentleman come-NPST-IND
'The boss comes/will come.'
- (6) mahattea a-aw-a
gentleman come-PST-IND
'The boss came.'
- (7) unnæhee hungak prəsiddə kene-k
3SG.M very famous person-NOM
'He is/was a very famous person.'

And there are many languages with (at least some) zero inflections but always an overt copula (translated as COP in example 9), such as Cambodian (Austroasiatic, Cambodia):

- (8) vi:ə tɻu phsa:r
he go market
'He goes/went/will go to market.'
- (9) mən-s nùh ci:ə kru:
man that COP teacher
'That man is a teacher.'

Instead, the universal governing occurrence of a copula is the semantic type of predication: if predicate adjectives (predication of a property) requires a copula in a language, then predicate nominals (predication of object class) requires a copula as well (Croft 1991:130; Stassen 1997:127).

2. How many languages in typological study?

It is impossible to examine all of the approximately six thousand languages of the world (Comrie, this volume) in order to describe their typological variation and infer language universals. As with other sciences, a sample of languages must be taken, and the universals that are inferred are dependent on the quality of the sample. The standard approach to sampling is to ensure that the languages in the sample are drawn from as many different language families and different geographical regions as possible given the

size of the sample (anything from a dozen to several hundred languages). The reason for this is that for statistical purposes, the grammatical constructions in the different languages in the sample should be historically independent. For example, the fact that the numeral precedes the noun in both English and German is probably due to their descent from a common ancestor language (proto-Germanic), hence including both languages in a typology of word order is to be avoided.

However, a proper sample must be large enough to capture all of the diversity of human language. With a large sample, the likelihood increases that the typological traits of two or more languages in the sample are historically related, either by descent from a common ancestor or through language contact, recent or ancient. In fact, many linguists believe that all modern languages may be descended from a single common ancestor, even if the time depth and amount of change from that time means we can never construct a complete language family tree with a high degree of confidence. Thus it is possible that some typological traits have been a result of the form of Proto-World, or at least some very ancient protolanguage. For example, the indigenous languages of Papua New Guinea that do not belong to the Austronesian language family are almost entirely subject-object-verb (SOV) in word order. The island of Papua New Guinea was colonized at least 35,000 years ago if not earlier, and it is likely that the SOV word order of all the languages on Papua New Guinea are historically related through common ancestry or contact for at least 35,000 years. Some linguists have even suggested that SOV word order may have originated with Proto-World, though the evidence for this claim is fragmentary and circumstantial.

From a traditional statistical point of view, the fact that even distantly related languages may share typological traits through historical relations is problematic. There is an alternative to the sampling strategy, however, which is precisely to take historical relatedness into consideration when constructing and evaluating language universals. Dryer's sampling method is the first in linguistics to do this (Dryer 1989). He does not discard any language from his sample, even if they closely related genetically or in the same geographical instead. Instead, he aggregates the data from closely related languages into *genera* (a genetic grouping of approximately the same time depth as a major subgroup of Indo-European such as Romance or Germanic), and then aggregates those results into six large linguistic areas in the world (Africa, Eurasia, Southeast Asia & Oceania, Australia & Papua New Guinea, North America and South America). For example, using this method, Dryer demonstrates that the hypothesized relationship between adjective-noun order and object-verb order is a consequence of a high preponderance of this correlation among genera in the Eurasian area, and is not an inherent causal connection between these two word orders (Dryer 1991).

More recently, Maslova and her colleagues have used a sophisticated mathematical model by taking pairs of closely related languages, inferring from the match or mismatch of typological traits in those pairs the likely rate of change of the typological trait, in order to determine what the stationary distribution of the typological traits are, and hence their inherent dominance or preference (Maslova 2000). For example, in an unpublished paper Maslova and Rakhilina demonstrate that the current distribution of alignment patterns between ergative and accusative (see §4) is not a stationary distribution: it

appears that a stationary distribution would imply a higher proportion of the accusative alignment pattern, and hence a higher dominance of the accusative pattern.

The best solution to the problem of historical relatedness of typological traits in crosslinguistic research is to use the linguistic family tree directly to model the evolution of typological traits through time and identify any causal relations between those traits. This method has not been applied to linguistic traits in published studies (but see unpublished work by Michael Dunn and Russell Gray), but it has been applied to cultural traits using language family trees. For example, Holden and Mace (2003) use the structure of the Bantu language family tree, derived from linguistic data using techniques from molecular phylogeny, to support a causal connection between the adoption of cattle and the loss of matrilineal descent patterns in prehistorical southern Africa. Pagel and Meade (2005) use a similarly constructed tree for Indo-European and argue that proto-Indo-European culture had monogamy and a dowry system, and daughter societies were first to polygyny and then to the absence of a dowry or presence of a bride-price (i.e. a shift to polygyny causes the abandonment of a dowry). The same methods can be applied to, for example, the relationship between object-verb order and genitive-noun order—two orders which are strongly though not completely correlated across languages—in various language families, in order to determine the causal connection between those orders and the directionality of causation.

These new techniques will allow us to use historical relations among languages to discover language universals rather than treating them as obstacles in inferring language universals. The techniques drawn from molecular phylogeny will have to be modified in order to incorporate historical relations through language contact as well as common ancestry. Finally, the integration of historical relations between languages in inferring language universals requires a shift to analyzing language structure as part of a dynamic evolving linguistic system over time, and language universals as universal constraints on the evolution of linguistic systems. This shift in approach, necessitated by methodological issues in constructing language samples, actually dovetails with the general shift in typological theory from universals of synchronic language systems to universals constraining language change (§6).

3. How does one person use data from so many languages?

A question that often comes to mind about typological research, especially outside the field of linguistics, is how can a single typologist or a small group of collaborators acquire the data from the large number of languages in a sample? It is impossible for a single linguist to know directly a hundred or a few hundred languages, or to consult with a similar number of native speaker consultants. For that reason, typological research is necessarily built on the foundations of sound documentation of the world's languages by native speakers or by field linguists and their native speaker consultants (Munro, this volume). Many typologists have also done fieldwork in particular languages, language groups or language areas. The severe endangerment of the vast majority of languages in the world has long been of great concern to typologists.

Typologists have used a variety of resources to obtain the data for typological research. Perhaps the most important is a reference grammar of a language. A surprisingly large percentage of the languages of the world have reference grammars, though they vary substantially in breadth and depth of coverage and in quality. A reference grammar is quite different from a pedagogical grammar: it is designed not for learning a language, but as a reference source for the major grammatical constructions of the language. A reference grammar is organized around families of constructions (and also the sound structure and morphological structure of words in the language), although this organization varies from grammar to grammar. A reference grammar presupposes to a greater or lesser extent a knowledge of basic linguistic terminology and a knowledge of the language itself (for example, many reference grammars give examples only in the script of the language).

In order to use data from a language that one does not speak or has not studied in detail, a linguist must be able to analyze the structure and the meaning of phrases or sentences in the language. The meaning is generally provided by a translation of an example into English (or French or Spanish, or whatever is the language that the reference grammar is written in). The structure of the example is ideally provided by a separate line by the author of the reference grammar, sometimes referred to as the *interlinear morpheme translation* (IMT). Examples of IMTs can be seen in examples (4)-(9) above. An example of a sentence from Mokilese, an Austronesian language of Micronesia, with an IMT is given in (10).

- (10) mine woaroa-n woal-o war
 exist CLF-3SG man-that canoe
 ‘That man has a canoe.’

The last line, the *gloss*, gives the meaning of the Mokilese original sentence as a whole: from the gloss, we can see that this is an example of the predication of possession, to use the technical linguistic description of the function of the sentence. The middle line, the IMT, glosses each morpheme of the original, including affixes and other meaningful internal word changes, following relatively widely used conventions. The IMT therefore provides the structural analysis of the sentence: the order of elements (e.g., the verb comes first in the sentence); which elements are independent words and which are not (e.g. ‘that’ is a suffix on ‘man’); and the presence of special grammatical elements (e.g. the possessive classifier for the man’s canoe, and the person-number inflectional suffix on the classifier). The IMT also shows how differently the function of predication of possession is expressed in the language: in this case, a more literal translation of the Mokilese sentence would be ‘That man’s canoe exists’.

The presence of IMTs allows a linguist to rapidly recognize the grammatical structure of the language in the description even if the linguist has not previously studied the language. IMTs are used widely in contemporary reference grammars. Many grammars, especially older ones, do not use IMTs, and therefore a linguist must study the reference grammar (and a dictionary if available) in greater detail in order to analyze the structure of the constructions of interest in the language.

Reference grammars vary significantly in the constructions that they cover, and the depth in which they cover them. A typologist may discover that some (or many) of the reference grammars do not describe the construction s/he is interested in. One may need to accept that there will be gaps in the data in the sample (or one will replace the language in the sample with another language possessing a grammar with the relevant description). More recently, typologists have developed questionnaires to obtain data for particular constructions which are typically poorly described in reference grammars. One of the first questionnaires used a set of 200 sentence contexts to elicit tense-aspect constructions; the questionnaire was distributed to native speakers or specialists in 65 different languages (Dahl 1985). More recently, experimental elicitation techniques have been used to avoid the problem in questionnaires of translation from the language of the questionnaire. For example, a set of 71 pictures of different spatial scenes was used to elicit constructions for spatial relations in nine diverse languages (Levinson et al. 2003), and a set of 61 video clips of different cutting and breaking scenes was used to elicit different verbs of cutting and breaking in 28 languages (Majid et al. 2007).

Questionnaires and experimental elicitation techniques have the advantage of allowing a linguist to directly elicit the constructions expressing the function s/he is investigating. However, designing questionnaires and elicitation strategies is extremely difficult. Also, questionnaires and elicitation, while valuable in themselves, can benefit from a broader description of the language than just the constructions under investigation. Many times a typologist discovers an unexpected correlation of constructions. For example, one might not consider the English conditional construction in (11) and the topic construction in (12) to be grammatically related:

- (11) If he comes, I will stay.
 (12) (As for) me, I will stay.

But in a number of languages, including Hua, a Papuan language of New Guinea, the antecedent of a conditional and a topic are marked grammatically in the same way (in the case of Hua, with the suffix *-ve*; Haiman 1978:570-71):

- (13) *e-si-ve* *baigu-e*
 come-3SG.FUT-*ve* will.stay-1SG
 ‘If he comes, I will stay.’
- (14) *dgai-mo-ve* *baigu-e*
 I.EMPH-CONN-*ve* will.stay-1SG
 ‘(As for) me, I will stay.’

This grammatical connection is at first surprising, although an explanation is forthcoming: the situation described by the antecedent of a conditional provides the framework for evaluating the consequent that follows it, and the referent described by a topic phrase constitutes the framework for the following discourse (Haiman 1978:577-86). This grammatical and semantic connection would probably not have been detected if one constructed a questionnaire for conditional constructions; but it might be observed if one consulted reference grammars for a wide range of languages, and noticed the

similarity of form. Thus reference grammars continue to have great value for typological research.

4. How can one compare grammatical structures from many different languages?

A fundamental characteristic of the typological approach to language is that one begins by comparing a wide range of languages for the grammatical construction in question. The most important observation that comes from that process is the extraordinary diversity of grammatical structures that are found. For example, we noted above that English expresses the predication of possession with a transitive verb *have* as in *That man has a canoe*, but the Mokilese expression is literally ‘That man’s canoe exists’, with an intransitive verb ‘exist’. Other constructions for the predication of possession would translate literally as ‘A canoe is located at that man’, ‘That man is with a canoe’, ‘For the canoe exists that man’, and ‘As for that man, a canoe exists’ (Heine 1997:92). Each of these constructions has its own distinctive grammatical structure, and in addition to these major types there are a number of other less common types.

What is the basis for crosslinguistic comparison? The tremendous diversity of grammatical structures point to an answer that was offered in the seminal paper in typology, Greenberg (1966), and has been reiterated since: meaning or function. For example, in the study of possession, Heine compares whatever grammatical constructions are used in a language to express the function of predication of possession. Possession is what is traditionally described as a semantic relation: an ownership relation between a possessor (usually a person) and a possessum (usually an artifact). Predication is generally described as a pragmatic function, having to do with the packaging of information in a sentence: the possession relation is asserted, in contrast to attributive possession (as in *that man’s canoe*), in which the possessive relation is used to modify the description of the referent *canoe*.

The use of function as a basis for crosslinguistic comparison allows one to avoid problems arising from the usage of traditional grammatical terminology. For example, in English a contrast is made between relative clauses with *finite* (fully inflected) verb forms, as in (15), and participles, a nonfinite verb form, as in (16):

- (15) the papaya that the ants ate
(16) the papaya eaten by the ants

A crosslinguistic study of “relative clauses” immediately encounters the situation that some languages such as Turkish (Altaic, Turkey) use participial or nominalized verb forms heavily, as in example (17) (Comrie 1989:142; see Keenan and Comrie 1977):

- (17) Hasan-in Sinan-a ver-diğ-i patates-i yedim
Hanan-GEN Sinan-DAT give-NR-his potato-ACC I.ate
‘I ate the potato that Hasan gave to Sinan.’

However, this is an artifact of the traditional grammatical terminology, which treats finite and nonfinite constructions completely differently. Keenan and Comrie define

relative clauses functionally, in terms of a proposition modifying a referent (the noun); using this definition, Turkish has constructions comparable to English relative clauses (and participles).

One widely held but invalid assumption about the crosslinguistic comparison of grammatical structures is that one can only compare grammatical structures by using the same grammatical categories across languages. Languages can vary to a remarkable degree in basic categories of grammar. For example, a plausible candidate for a pair of universal grammatical categories are the categories of subject and object of a verb:

(18) The woman didn't run.

(19) The snake bit the man.

The sentence in (18) has only a single phrase (*the woman*) referring to a participant in the event denoted by the verb (*run*). Such a sentence is intransitive, and *the woman* is labeled the 'subject'; we will refer to intransitive 'subject' with the label S. The transitive sentence in (19) on the other hand has two phrases referring to the two participants in the event (*bit*). It seems completely natural, indeed even necessary, that the first phrase, *the snake* (labeled A, mnemonic for 'agent') should belong to the same grammatical category as *the woman* in (18). Both *the woman* and *the snake* occur before the verb. Substitution of a pronoun for *the woman* would require the subject form *she*, not *her*. The grammatical category grouping S and A would be called 'subject'. The second phrase in (19), *the man* (labeled P, mnemonic for 'patient') is grammatically different. It occurs after the verb, and substitution of a pronoun for *the man* in (19) would require the object form *him*, not *he*. The grammatical category consisting of P is generally called the 'object'.

But many languages do not categorize the phrases referring to the participants in events in the same way. Compare the translations of (18) and (19) in Yuwaalaraay, an Aboriginal language of Australia:

(20) wa:l ɲama yinar-∅ banaga-ɲi
 not that woman-ABS run-NFUT
 'The woman [S] didn't run.'

(21) ɖuyu-gu ɲama ɖayn-∅ yi:-y
 snake-ERG that man-ABS bit-NFUT
 'The snake [A] bit the man [P].'

Yuwaalaraay does not have subject and object in the English sense. The grammar of participants is expressed by case suffixes on the noun. In an intransitive sentence like (20), the phrase labeled S has no suffix (notated here with the zero symbol -∅). In a transitive sentence like (21) however, what an English speaker would call the "subject", A, has a case suffix *-gu*, which is called the *ergative* case (abbreviated ERG), and the "object" phrase P has no suffix, like the "subject" S in (20). In other words, whereas English categorizes both A and S together (as subject) and distinguishes P (as object), Yuwaalaraay categorizes P and S together (called the *absolutive*, abbreviated ABS) and distinguishes A (as the ergative).

This difference between English (and many other languages) on the one hand, and Yuwaalaraay (and many other languages) on the other, is very striking. It seems very unnatural to us to group together S and P against A—subject and object in the English sense seem to be such basic categories of grammar. Since the terms ‘subject’ and ‘object’ are so loaded with theoretical and traditional connotations, most typologists discard them in describing the contrast between languages like English and languages like Yuwaalaraay, and instead use the terms *nominative* and *accusative*, from the Latin case terms, for the English categories. The crosslinguistic difference between English and Yuwaalaraay is called a difference in *alignment*.

The difference between English and Yuwaalaraay grammatical patterns here seems to make the two languages entirely different, so that no comparison, let alone universals of language, can be derived from the study of alignment systems. But this is not true. Crosslinguistic comparability of grammatical relations is based on the semantic roles, or more precisely prototype clusters of semantic roles, represented by S, A and P: one can compare how the agent of ‘bite’ and the single participant of ‘run’ are encoded grammatically across languages. More strikingly, there are language universals that can be stated no matter what the alignment system of the language is. For example, the absolutive is zero-coded in Yuwaalaraay and many other languages, in contrast to the overt coding of the ergative case. In languages with a nominative-accusative alignment and case inflections, the nominative is usually zero-coded in contrast to overt coding for the accusative. These are not the only types found: the universals for each alignment system can be more precisely formulated to cover all types, as in (22a-b):

- (22) a. The accusative is encoded with at least as many morphemes as the nominative.
b. The ergative is encoded with at least as many morphemes as the absolutive.

The universals for the two alignment systems in (22a-b) have a single explanation. Noun phrases with nominative referents are more frequent in discourse than noun phrases with accusative referents, and noun phrases with absolutive referents are more frequent in discourse than noun phrases with ergative referents. A far-reaching universal encompassing (22a-b) and many other grammatical categories is: categories of meanings that are more frequent in discourse are likely to be encoded with fewer morphemes (Croft 2003; see §6). Hence it does not matter what the grammatical categories of a language are in order to find language universals. One must simply identify how concepts are encoded in a language. The language universals are based on how concepts are encoded, not a set of universal grammatical categories.

The methodological necessity of comparing constructions encoding the same functions across languages matches the functionalist orientation of most typologists. In order to discover language universals, a typologist compares how function is encoded in linguistic form. A crosslinguistic perspective does not allow a typologist to analyze form autonomously from its function. The very fact of crosslinguistic variation in the formal encoding of linguistic function, however, implies that linguistic form is separate from linguistic function, and that linguistic form is at least partly arbitrary—otherwise, all languages would have the same grammatical structure.

(25) Jamiltepec Mixtec (Mixtecan, Mexico)

a. vēhē	lúhlu	b. uvi	vēhē
house	little	two	house(s)
Noun	Adj	Num	Noun

On the other hand, languages with the adjective preceding and numeral following are virtually unattested (although there are a few; the existence of exceptions will be discussed below).

The pattern of attested vs. unattested (or at least extremely rare) language types can be given in the four-cell table below:

	<i>Noun-Adjective order</i>	<i>Adjective-Noun order</i>
<i>Numeral-Noun order</i>	attested (Jamiltepec Mixtec)	attested (English)
<i>Noun-Numeral order</i>	attested (Kosraean)	extremely rare

Table 1. Attested vs unattested adjective and numeral word orders.

The generalization can itself be described in terms of an IMPLICATIONAL UNIVERSAL:

(26) If a language has Adjective-Noun word order, then it (almost always) has Numeral-Noun word order.

The discovery of implicational universals of word order by Greenberg (1966) demonstrated that there exist universal properties of language that do not require all languages to be identical in specific properties. The implicational universal in (26) is not by itself a description of a fact about the grammar of a particular language. In fact, one could not even identify the implicational universal without looking across a set of languages. The implicational universal captures a contingent relationship between Adjective-Noun order and Numeral-Noun order. Nevertheless, this contingent relationship must be a part of individual speakers' knowledge of language structure and meaning. In particular, a language will not (or is extremely unlikely to) emerge or change to a type that has both noun-numeral order and adjective-noun order.

Hundreds of implicational universals have already been discovered, and more are discovered every time a typologist investigates a new area of grammar. The existence of so many implicational universals requires a rethinking of the nature of Universal Grammar, which is usually thought of as a set of unrestricted universals. The part of Universal Grammar that consists of unrestricted universals specifying ways in which all languages are identical captures only a very small portion of what is universal about language. It misses most of what is universal about language beyond the basic design features. The presence of large numbers of implicational universals requires a model

which allows for a great deal of variation in grammatical structure across languages, but constrains that variation to a significant degree in many different dimensions. This basic observation about language universals has not been addressed by most syntactic theories.

The second general conclusion about language universals that is revealed by typological research is that the constraints on language variation that are represented by implicational universals are not exceptionless in the way that the design features of language are. In the example of the relationship between numeral-noun order and adjective-noun order given above, we noted that the fourth language type, noun-numeral order and adjective-noun order, is not actually unattested: a small number of languages of this type do exist. This is almost always observed for language universals, especially with advent of larger and larger language samples.

This conclusion also changes how language universals must be understood. One cannot conceive of language universals (beyond the design features) as specifying what constitutes a possible human language. The noun-numeral & adjective-noun language type is not impossible; a few such languages exist. But it is far less frequent than the other three types. Moreover, the other language type with modifiers on the opposite side of the noun, adjective-noun and noun-numeral, while frequent, is not as frequent as the types in which both modifiers precede or both modifiers follow. These differences in the likelihood of language types are significant and must be explained by linguistic theory. In other words, typology shifts the scientific question about language universals from “What is a possible language type, and why?” to “What is a more probable language type, and why?”

For this reason, typologists have turned to more sophisticated quantitative methods in order to identify valid language universals. If all types of languages exist in the area of grammar under investigation, then one must be confident that differences in likelihood are statistically valid. The increase in quantitative sophistication in language sampling was discussed in §2. In the case of implicational universals, Maslova (2003) gives statistical tests for identifying valid implicational universals when one or more types in a table such as Table 1 is of much lower frequency than the other types.

An important type of model for inferring the language universals underlying generalizations such as the implicational relation in (26) is the *competing motivations* model. A competing motivations model posits two or more factors that determine language structure. However, the motivations typically do not determine a single grammatical pattern because they are often in conflict. In the case of conflict, there is no single optimal grammatical pattern that satisfies all of the competing motivations, and instead one finds cross-linguistic variation over several suboptimal patterns. In this way, universal properties of language (the motivations) give rise to cross-linguistic diversity.

For example, Greenberg proposed two competing motivations for implicational universals of word order. The first, *dominance*, can be thought of as simply a default preference for one order over another. For example, noun-adjective order (NA) is dominant, as is numeral-noun order (NumN). The second, *harmony*, can be thought of as

a dependent relation of one word order upon another. For example, AN order is harmonic with NumN order and NA order is harmonic with NNum order.

Greenberg's two motivations compete with each other, and the result is described in the following principle:

- (27) A dominant order may occur at any time, but a recessive order occurs only when a harmonic order is also present.

The principle in (27) accounts for the distribution of languages in Table 1. The upper left cell is the language type with both dominant orders (NA and NumN), which are not harmonic with each other. The other two attested types have one recessive order, but the harmonic order is also present. The extremely rare type would have both recessive orders (AN and NNum), neither of which is dominant. That is, the extremely rare type is not motivated by either dominance or harmony, which accounts for its rarity. Finally, one cannot satisfy both motivations at once, since the dominant orders are not harmonic with each other.

A competing motivations model can be found by inspection with a simple case such as the two interacting word orders in Table 1. But it turns out that comparing the word orders of several different constructions across languages is much more complex: most possible types are attested, but at highly varying frequencies. Justeson and Stephens (1990) use a statistical technique, log-linear analysis, to examine the relationships between multiple word orders. Log-linear analysis allows one to construct a model of the simultaneous interactions between many word orders; the model that best fits the data provides the best model of which word orders actually might be causally connected. The best fit model for the word orders investigated by Justeson and Stephens identify the dominance and harmonic relations that are actually supported by the data in their 147-language sample. Reassuringly, Justeson and Stephens arrive at the same result that Dryer came to, namely that there is no direct causal relation between adjective-noun order and object-verb order.

Another example of the use of sophisticated statistical techniques in typology is found with an important recent development in modeling language universals, the *semantic map* model. The semantic map model allows the typologist to identify language universals without assuming that grammatical categories are the same across languages, as we observed in §4. There, it was pointed out that some languages divide the semantic roles in transitive and intransitive verbs (A, P and S) in a nominative-accusative pattern (A+S vs. P) while others divide them in an ergative-absolutive pattern (A vs. S+P). Still other languages use a tripartite pattern, distinguishing all three of A, S and P, while others have a neutral pattern, not distinguishing any of them. No language (or virtually no language) groups together A and S and distinguishes that group from P.

The range of variation and the limitation of attested types of grammatical relations can be represented by a *conceptual space* linking A to S and P to S, as in Figure 1:

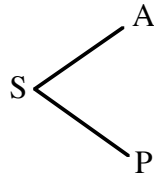


Figure 1. Conceptual space for semantic roles.

The attested language types can be mapped as connected regions on this three-point space:

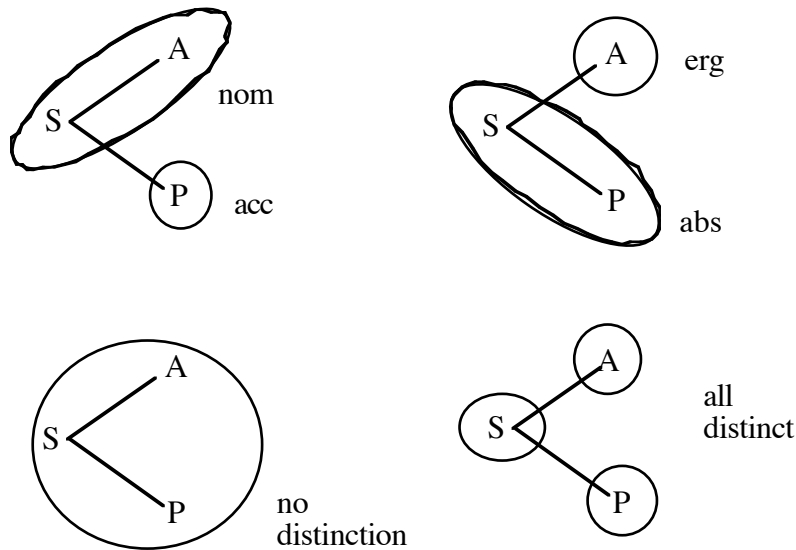


Figure 2. Map of attested systems of grammatical relations

The unattested (or extremely rare) type of A+P vs. S cannot be mapped on the conceptual space in Figure 1: A cannot be grouped with P without also including S, according to the connections in Figure 1. Hence the conceptual space represents what is universal about the relationships among A, S and P, while allowing for the variation of attested language types illustrated by the semantic maps in Figure 2.

The semantic map model is a powerful tool for identifying language universals and separating language universals from the arbitrary aspects of crosslinguistic variation, and has been widely used. However, it is impractical for identifying the conceptual space when that space consists of many more points than the three exemplified in Figures 1-2, or when the range and frequency distribution of attested language types is complex. However, these problems can be solved by using multidimensional scaling and related multivariate techniques. For example, multidimensional scaling was applied to the results of the experimental elicitation for the 71 pictures of spatial relations described in §3, allowing typologists to identify the most significant semantic dimensions of spatial relations in determining attested systems of spatial relations (prepositions, postpositions and case inflection; Levinson et al. 2003; Croft and Poole 2008).

The richness, diversity and complexity of typological evidence are daunting to analyze. Typology has only recently begun to use more sophisticated quantitative tools to allow researchers to go beyond the easy to identify language universals to the universals that are harder to detect through the noise of arbitrary crosslinguistic variation. The initial successes in using such tools indicates that typologists will be able to find many more restricted universals of language.

6. Explanations for language universals

Most typologists seek functional explanations for language universals. This is chiefly because the basis of crosslinguistic comparison is how function is encoded in grammatical form. Nevertheless, there are many aspects of language function, and different aspects have been proposed to explain different kinds of language universals.

Proposed explanations for universals of word order (and also the order of affixes) fall into two general categories: language processing in production and comprehension, and diachronic explanations. Most models of word order universals include harmonic patterns, in which two or more word orders are correlated. The processing explanation is based on theories of how easy or difficult it is for speakers construct or parse utterances: put crudely, the word order combinations that facilitate parsing are more common crosslinguistically, and those that make parsing more difficult are less common (a detailed theory is presented in Hawkins (1994, 2004)). A diachronic explanation of word order correlations is based on the fact that the actual grammatical constructions used for different word orders are historically related. For example, there is a very tight harmonic correlation between genitive-noun order and the order of adposition and noun: genitive-noun correlates with postpositions, and noun-genitive with prepositions (Greenberg 1966). The most likely explanation for this correlation is that adposition constructions are very frequently historically derived from genitive constructions via the process of grammaticalization (Greenberg 1969; Lehmann 1982; Bybee et al. 1994; Hopper and Traugott 2003). This can be observed in the history of English, where the prepositional phrase *inside the house* derives historically from the noun-genitive construction *in the side of the house*. Grammaticalization theory has uncovered a large number of universals of language change, specifically processes by which words in particular syntactic constructions evolve into grammatical elements.

The semantic map model also lends itself to a functional explanation for patterns of the expression of grammatical relations (see Figure 2 in §5), spatial relations and many other grammatical categories and constructions. The explanation is based on a universal conceptual space. The conceptual space represents the degree of conceptual relatedness of the situations or concepts represented by the points in the conceptual space. The structure of the conceptual space is hypothesized to represent general properties of human cognition and conceptualization. The language-specific categories represented by the semantic maps are partly arbitrary (hence the crosslinguistic variation), but they must conform to the constraints imposed by the structure of the conceptual space (Croft and Poole 2008).

Economic motivation is extremely common crosslinguistically: many universals of grammatical inflection are explainable in terms of frequency differences, and economy is increasingly being invoked to explain universals of syntactic constructions as well (Bybee 2006). Iconic motivation is also extremely common. In fact, iconicity is often taken for granted. For example, the fact that referents and their modifying properties are syntactically constituents (e.g., in *Two boys ate five pizzas* the number of boys is two, not five) is basically assumed, but it represents the iconic motivation that conceptual relations support syntactic constituency. Also, most syntactic theories contain principles specifying that the syntactic arguments of a verb must match the semantic participants in the event denoted by the verb in number and type; this is also an iconic principle. Economy and iconicity frequently are in competition, leading to the crosslinguistic variation of the type observed in Table 2.

The competition between processing mechanisms of different kinds and the different types of conceptual relations holding among components of the situation expressed by a grammatical construction leads naturally to a dynamic model of language and language universals. Many typologists have proposed evolutionary models of how language structures adapt to the functions they perform in communication and the constraints on comprehension and production of utterances in language use (e.g. Greenberg 1979; Croft 2000, 2003; Givón 2002). In an evolutionary approach, a linguistic system is conceptualized as a stage in the process of language change, which can shift by the choices of speakers. Language universals are the result of universal cognitive and social interactional forces that shape speakers' choices at all timescales, from the immediate discourse situation to the lifetime of a speaker to the transmission of language across generations. The evolutionary approach to grammar emerging from typology its integration with related approaches, including sociolinguistics (Coulmas, this volume) and the usage-based model (Bybee 2006), promises a major step forward in our understanding of language.

References

- Baker, Mark. 2003. *Lexical categories: verbs, nouns and adjectives*. Cambridge: Cambridge University Press.
- Bybee, Joan L. 2006. *Frequency of use and the organization of language*. Oxford: Oxford University Press.
- Bybee, Joan L., Revere D. Perkins & William Pagliuca. 1994. *The evolution of grammar: tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.
- Chomsky, Noam. 1976. *Reflections on language*. London: Fontana.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
- Comrie, Bernard. 1989. *Language universals and linguistic typology* (2nd ed.). Chicago: University of Chicago Press.
- Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: University of Chicago Press.
- Croft, William. 2000. *Explaining language change: an evolutionary approach*. Harlow, Essex: Longman.

- Croft, William. 2003. *Typology and universals*, 2nd edition. Cambridge: Cambridge University Press.
- Croft, William. 2008. Methods for finding language universals in syntax. *Universals of language today*, ed. Sergio Scalise, Elisabetta Magni and Antonietta Bisetto, 145-64. Berlin: Springer.
- Croft, William and Keith T. Poole. 2008. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics* 34.1-37.
- Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Basil Blackwell.
- Dik, Simon C. 1997. *The theory of functional grammar*, ed. Kees Hengeveld (2 vol.). Berlin: Mouton de Gruyter.
- Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13.257-92.
- Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68:81-138.
- Givón, Talmy. 2002. *Bio-linguistics: the Santa Barbara lectures*. Amsterdam: John Benjamins.
- Greenberg, Joseph H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Grammar*, ed. Joseph H. Greenberg, 2nd edition, 73-113. Cambridge, Mass: MIT Press. (Reprinted in Greenberg 1990, 40-70.)
- Greenberg, Joseph H. 1969. Some methods of dynamic comparison in linguistics. *Substance and Structure of Language*, ed. Jan Puhvel, 147-203. Berkeley and Los Angeles: University of California Press. (Reprinted in Greenberg 1990, 71-118. Stanford: Stanford University Press.)
- Greenberg, Joseph H. 1979. Rethinking linguistics diachronically. *Language* 55.275-290. (Reprinted in Greenberg 1990, 673-91.)
- Greenberg, Joseph H. 1990. *On language: selected writings of Joseph H. Greenberg*, ed. Keith Denning & Suzanne Kemmer. Stanford: Stanford University Press.
- Haiman, John. 1978. Conditionals are topics. *Language* 54.564-589.
- Haiman, John. 1983. Iconic and economic motivation. *Language* 59.781-819.
- Haiman, John. 1985. *Natural syntax: iconicity and erosion*. Cambridge: Cambridge University Press.
- Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
- Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
- Heine, Bernd. 1997. *Possession: cognitive sources, forces, and grammaticalization*. Cambridge: Cambridge University Press.
- Holden, Clare Janaki and Ruth Mace. 2003. Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proceedings of the Royal Society of London Series B* 270.2425-33.
- Hopper, Paul & Elizabeth Traugott. 2003. *Grammaticalization* (2nd edition). Cambridge: Cambridge University Press.
- Joos, Martin (ed.). 1957. *Readings in Linguistics I*. Chicago: University of Chicago Press.

- Justeson, John S. & Laurence D. Stephens. 1990. Explanations for word order universals: a log-linear analysis. *Proceedings of the XIV International Congress of Linguists, vol. III*, ed. Werner Bahner, Joachim Schildt and Dieter Viehweger, 2372-76. Berlin: Mouton de Gruyter.
- Keenan, Edward L. & Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8:63-99.
- Lehmann, Christian. 1982. *Thoughts on grammaticalization: a programmatic sketch, Vol. I*, Arbeiten des Kölner Universalien-Projekts, 48, Köln: Institut für Sprachwissenschaft. (Revised edition published by LINCOM Europa, München, 1995. Revised edition reprinted as Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt, 9. Erfurt: Seminar für Sprachwissenschaft der Universität.)
- Levinson, Stephen C., Sérgio Meira, and the Language and Cognition Group. 2003. 'Natural concepts' in the spatial topological domain—adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language* 79.485-516.
- Majid, Asifa, Melissa Bowerman, Miriam van Staden and James S. Boster. 2007. The semantic categories of cutting and breaking events: a crosslinguistic perspective. *Cognitive Linguistics* 18.133-52.
- Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4.307-33.
- Maslova, Elena. 2003. A case for implicational universals. *Linguistic Typology* 7.101-108.
- Pagel, Mark and Andrew Meade. 2005. Bayesian estimation of correlated evolution across cultures: a case study of marriage systems and wealth transfer at marriage. *The evolution of cultural diversity: a phylogenetic approach*, ed. Ruth Mace, Clare J. Holden and Stephen Shennan 235-56. London: UCL Press.
- Stassen, Leon. 1997. *Intransitive predication*. Oxford: Oxford University Press.
- Van Valin, Robert D. Jr. & Randy J. LaPolla. 1997. *Syntax: structure, meaning and function*. Cambridge: Cambridge University Press.

Appendix: abbreviations used in the examples

1	first person
3	third person
ABS	absolutive
ACC	accusative
CLF	classifier
CONN	connective
COP	copula
DAT	dative
EMPH	emphatic
ERG	ergative
FUT	future
GEN	genitive
IND	indicative
M	masculine
NFUT	nonfuture
NOM	nominative
NPST	nonpast
NR	nominalizer
PL	plural
PST	past
SG	singular