Linguistic Typology
Anders Holmberg

1.   Introduction[1]

Linguistic typology, as usually conceived, is a research programme combining certain aims and objectives with certain methods and a preferred theoretical framework. The aim is to understand linguistic variation. More specifically, the aim is to distinguish between properties which are shared across languages for historical reasons (either shared parenthood or language contact), and for other reasons, to do with 'the nature of language' in some sense, and ultimately understand what those other reasons are. The preferred method is large scale comparisons of as many languages as possible, sampled so as to control for genealogical and areal biases. The outcome is a typology or typologies of languages along a variety of parameters such as word order, stress placement, vowel systems, case systems, possessive constructions, etc., etc.   The theoretical challenge is to explain the cross-linguistic generalizations which are not due to historical contingencies. The preferred mode of explanation is in terms of functional rather than formal notions.

This paper will describe the basic properties of the research programme, some of its strengths and weaknesses, and some recent trends. It is written from the perspective of an outsider (a generative linguist) but a mainly sympathetic outsider interested in promoting closer collaboration between typological and generative linguistic research.

The driving force behind typological research is the belief that there are systematic similarities among languages, leading to patterns of variation, which are not due to historical contingencies but to 'the nature of language', in some sense, and that these patterns and shared properties can be detected and mapped on the basis of careful comparison of languages, and explained at least in part on the basis of functional considerations, mainly to do with efficiency of processing and communication. The research program has undergone some significant development in the last twenty years, where the new elements are: A much better picture of areal effects on the form of languages (so much so that linguistic typology and linguistic geography can now be seen as facets of the same research program), and a move away from the search for Greenbergian implicational universals towards a purely descriptive enterprise. I will comment on these developments below.

Linguistic typological research can point to some indisputable and quite astounding results. As a result of this research we now have detailed knowledge of the distribution of a very wide range of grammatical/linguistic properties, over the full range of languages described in some cases, over large subsets of languages in other cases. The most concrete result of this research is the World Atlas of Linguistic Structures (WALS) a database which is available as a book (Haspelmath & al. 2005) and online (the most recent edition at the time of writing is Haspelmath & al. 2011). It contains and systematises a huge amount of facts about 144 grammatical properties (called 'features'), in samples of the world's languages of varying sizes (up to 1500 languages for the largest samples), and including data from all in all 2,650 languages at the time of writing. WALS is characterized by Baker (2010) as "wonderful and frustrating". I will comment on both aspects.

2.   History: The Greenbergian paradigm

Modern linguistic typological research starts with Joseph Greenberg's paper 'Some universals of grammar with particular reference to the order of meaningful elements' (Greenberg 1963). Linguistic typology existed before this paper; in particular, there is a body of work on morphological types, i.e. synthetic vs. analytic languages and varieties thereof, going back to the 19th century (see Greenberg (1974), Croft (1999, 2003: 39-42)). However, Greenberg's (1963) paper very clearly defined a new research programme in linguistic research. The paper has all the major components of most subsequent research on linguistic typology: the global, comparative perspective and the aim to discover cross-linguistic generalisations which are not the result of genetic relatedness or language contact, which requires selection of a representative and balanced sample of languages for comparison. Greenberg's paper also exemplifies the 'statistical' approach which is typical of most subsequent typological work: the generalizations are not required to be absolute. Finally there is the ambition to get beyond the generalisations to a theoretical explanation in terms of more fundamental properties of language.

What Greenberg did was put forward a set of 45 cross-linguistic generalizations, termed 'universals', based on (primarily) a sample of 30 languages, selected with a view to get a certain spread in terms of genetic groupings. The universals include the famous word order universals, stating correlations between the ordering of the lexical categories and their complements, including Universals 3, 4, and 5 below.

Universal 3:   Languages with dominant VSO order are always prepositional.

Universal 4:   With overwhelmingly greater than chance frequency, languages with SOV order are postpositional.

Universal 5:   If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun.

The other universals concern morphology, including Universals 27 and 34.

Universal 27:   If a language is exclusively suffixing, it is postpositional, if it is exclusively prefixing, it is prepositional.

Universal 34:   No language has a trial number unless it has a dual. No language has a dual unless it has a plural.

It is obvious now, and it was obvious to Greenberg at the time, that the sample of languages is not a balanced sample representing all the families and regions in the world in a principled fashion (for instance, as many as six of the languages are Indo-European), but the ambition is nevertheless present; thus all the continents (Africa, the Americas, Eurasia, and Australia/Oceania) are represented.

At the end of the paper Greenberg also presents, quite tentatively, a theoretical explanation of a subset of the universals, namely the word order universals. The explanation is in terms of the notions 'dominance' and 'harmony'. This also set the tone for much subsequent work in linguistic typology and, a bit later, generative grammar.

There were two interrelated aspects of Greenberg's work which, in particular, made it such an inspiration for linguistic research in typology as well as, later, in generative linguistics. One is the formulation of cross-linguistic implicational generalizations: If a language has property P, it also has

property Q. Even though, as noted by Haspelmath (2008), implicational statements are implicit in basically all typological work ('A language which has agglutinating verbal morphology, has agglutinating nominal morphology'), Greenberg formulated them as a set of explicit, interrelated, and, as it looked, easily testable hypotheses. The second is the fact that a set of the generalizations, in particular the word order universals, seemed to hang together, centering on the order of S, O, and V. This strongly suggested that languages form 'holistic types' of a sort that had not been observed before. There was unity, simplicity and some sort of rationality underlying the seemingly unconstrained surface variation.

There was a flurry of works in the late 60s and 70s taking Greenberg (1963) as a starting point, extending the programme to other languages , proposing explanations of the universals, and applying them to diachronic linguistics (Lehmann 1973, 1978, Venneman 1974, 1984, Hawkins 1979, 1983, Greenberg & al. 1978). In generative grammar the importance of Greenberg's universals became apparent when X-bar theory was generalized to all categories, making the phrase structure rule component redundant (Stowell 1981, Chomsky 1982, 1986), along with the introduction of the parametric theory of linguistic variation (Chomsky 1981). This made possible a view of the cross-linguistic word order universals as effects of the setting of the head-complement parameter, together with certain other auxiliary parameters (Travis 1984, Koopman 1984).

A proper extension and re-evaluation of Greenberg's work on the word order universals finally came with Dryer (1992), putting the project on a more secure footing both in terms of the number of languages (625 instead of 30), and the sampling of languages, and in terms of categories being considered. Dryer showed that some of Greenberg's universals do hold up (as statistical universals), under these more stringent conditions, others do not, but are shown to be areal effects, not detectable in Greenberg's small sample. As will be discussed below, within linguistic typology today, interest in research on word order universals has faded, to a significant extent.

3.   Explaining typological generalizations

There are basically two lines of explanation. One is the functionalist line, according to which cross-linguistic generalisations which are not historically based are (mainly) the result of a variety of functional factors to do with efficient processing and efficient communication, broadly speaking. Language has evolved to be as efficient as possible for communication (see for example Comrie 1989: 124, Hawkins 1994, 2004, Haspelmath 2008: 98ff, Newmeyer 1999: 105ff). [2] The other is a formalist line, according to which the cross-linguistic patterns are due to universal formal properties of language, at least some of which are ultimately biologically based, the result of genetically determined properties of the human language faculty. See Newmeyer 1999: 95-164 for a critical review of internal and external (mostly function-based) explanations in linguistics.

There has always been a strong tendency within research on linguistic typology to gravitate towards functional explanations rather than explanations in terms of innate universals, in fact, so much so that linguistic typology has been seen as almost synonymous with linguistic functionalism.[3] My own view on this is that there is no logically necessary connection between typology and functionally based explanation, but there is a practical one to do with the methodology: the

---

[2] There is also a cognitive-linguistic approach along the lines of of Langacker (1987), a variety of the functionalist approach which develops the idea that linguistic structure more or less directly reflects, or represents, conceptual structure. See Croft (2001: 108ff. and passim).

[3] See Croft (1995), who argues that there is a logical connection between typology and functionalism; see Newmeyer (1998: 348-349) for a rebuttal.

favoured method in linguistic typological research is comparative surveys on a large-scale in order to cover as much as possible of the existing variation and in order to establish, as far as possible, valid global generalisations. This has meant that the grammatical properties that are investigated/compared, are by necessity all easily observable 'surfacy' properties of the kind which are recorded even in sketchy descriptive grammars. One result of this is that the generalizations discovered have been probabilistic , riddled with exceptions, rather than absolute, because surfacy properties are subject to unpredictable variation to a greater extent than more abstract properties (as will be discussed below). This disfavours explanations in terms of universal, genetically determined properties of the language faculty, and favours explanations in terms of 'functional pressure', which are expected to allow for exceptions.

In generative grammar the aim is to uncover the universal properties of the human language faculty, including universal properties which are purely formal and ultimately based in the human genome. Such universal properties are typically not encoded in surface form, and the focus has therefore traditionally been on more abstract structural properties, which often can't be observed directly, but typically rely on negative evidence, i.e. require making distinctions between grammatical and ungrammatical sentences, which in turn requires extensive access to native speakers' intuitions (ideally the researcher is himself a native speaker of the language). This sort of data is seldom found in descriptive grammars, and in part for this reason, there has been only limited interest in typological research among generative grammarians.

More recently there has been a certain rapprochement between the functionalist and the biological approaches, from the formalist side. Within the Minimalist Program the prevalent view now is that there are few genetically determined properties that are specific to language. Instead, the form of language is largely determined by more general constraints and conditions on computational systems, even including some very general conditions on 'systems design' in the natural world (see Hauser, Chomsky, Fitch 2000, Chomsky 2005, Piatelli-Palmarini & Uriagereka 2008). This approach to UG is likely to lead to more openness towards functional explanations within generative grammar.[4]

 4.   **On the methodology of typological research**
The task of linguistic typology is to investigate linguistic variation. It should observe and record existing variation and establish valid generalizations about the variation, among the languages of the world. What is important, therefore, is not the number of languages investigated per se, but the variety: The investigation should include as wide a variety of languages as possible. The sampling of the languages is therefore crucial. Given the aim of discovering generalizations or patterns which are not the result of genealogical relationship or the result of extensive language contact, the set of languages compared should include languages, preferably in equal proportions, from every language family, down to as fine-grained a genealogical classification as possible, and from every corner of the world. The gnealogical class which is now (following Dryer 1989, 1992) commonly the basis for global comparison is *genus*, denoting a division at a time depth of no more than 3500-4000 years. The standard subfamilies of Indo-European (Germanic, Slavic, Celtic, etc.) would be genera, although Dryer (2011a) adds that "Celtic is perhaps a clearer example than Germanic or Slavic, both of which have a time depth considerably less than 3500 years". Thus a survey purporting to represent the

---

[4] This is not to imply that functional explanations would not earlier have been recognized within generative grammar; see Newmeyer (1998: 154-157). But it is fair to say that there has not been much interest in developing this approach, except within Optimality Theory; see Haspelmath (2008: 87-92) for discussion.

languages of the world should ideally include at least one representative from every genus of every language family. Although one may ensure an even distribution by including exactly one language from each genus, an alternative is to include any number of languages from all genera, but to count only genera (adopting some principle to decide which property is representative of the genus in cases where they are not consistent); see Dryer (1988, 1992), Rijkhoff & Bakker (1998).

In order to control for effects of extensive language contact, the sample should include representatives, in roughly equal proportions, from all over the world, i.e. from every continent, and every region of every continent, defined in some principled fashion. The importance of this methodological principle was amply demonstrated by Dryer's work on noun-adjective order (Dryer 1988), and has been confirmed by much work ever since (see Nichols 1990, 1992). Greenberg (1963) had proposed that AdjN correlates with OV, and NAdj with VO (Greenberg 1963: 100).[5] What Dryer found, when applying a principled sampling method, was that OV correlated strongly with AdjN order in Eurasia, while OV correlated, just as strongly, with NAdj in languages outside Eurasia. "In short, the previously believed tendency for OV languages to be AdjN is simply an Asian areal phenomenon /.../." (Dryer 1988: 188). Correspondingly, there was no global correlation between VO and NAdj order, but again there were areal effects, as in all the African language families SVO correlated with NAdj order ("an instance of an apparent pan-African tendency to place modifiers after the noun regardless of the order of verb and object": Dryer 1988:xx), while in the case of VSO order, the languages/families where NAdj was the dominant word order were all in North-West America. Apparently noun-adjective order is a property which is sensitive to language contact, leading to regional convergence.

One of the most interesting developments of linguistic typological research in the last twenty years or so is, indeed, the discovery that shared grammatical properties can be spread over very large areas, crossing genetic language boundaries, so that they cannot be explained by common ancestry, but still geographically restricted, so that they cannot be explained by universal grammar, leaving extensive language contact as the only possible explanation. This is the topic of the next section.

5.   Areal features

It has long been known that there are areal effects in linguistic variation, such that certain grammatical properties are shared by genetically unrelated or only distantly related languages in a region, the so called *Sprachbund* phenomenon. A famous case is that of the Balkan languages (including South Slavic languages, Romanian, Albanian, and Greek, all Indo-European, but only distantly related within this family) which share a number of features, including a suffixed definite article and loss of infinitive. Another well-known case is the Indian subcontinent, where the genetically unrelated Indo-Aryan and Dravidian languages have a number of properties in common, including retroflex stops and SOV order. The explanation for the *Sprachbund* phenomenon is, by all accounts, extensive language contact; people have intermingled within these regions,

---

[5] Greenberg (1963) proposed two generalizations, neither of which entails a direct correlation between object-verb and adjective-noun order. One is Universal 5 (see text above). The other is Universal 17: "With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun." In addition, Greenberg claimed that there was a general (universal) tendency for NAdj order, complicating the picture. See Dryer (1988) for discussion.

bilingualism/multilingualism and language shift have been widespread, resulting in the sharing of features, but where the social and political situation has nevertheless permitted languages to survive and continue to evolve as different languages.

One of the most interesting findings of typological research based on large scale global comparison, which could not have been found out with any other method, is the distribution of linguistic features over very large areas, in some cases continent-wide, or even spanning continents, and, importantly, in many cases spanning several language families, thus basically ruling out the possibility of a phylogenetic explanation (Nichols 1992). Above I mentioned Dryer's discovery that adjective-noun order is subject to continent-wide areal effects. Dryer (1998) shows that SOV order is not just an areal feature of the Indian subcontinent, but an Asian phenomenon more generally: Apart from Europe and South-East Asia, which both have strong VO preponderance, the Asian continent is solidly (S)OV.

 It is striking, when comparing the maps in WALS showing the distribution of grammatical features, how many of them show at least some areal effects which do not follow phylogenetic lines. The following are a few more examples:

Tone (Maddieson 2011). 527 investigated languages are divided into three types with respect to (lexical) tone:[6] Languages with no tones (307), with a simple tone system (132) and a complex tone system (88) (Maddieson 2011). The distribution is striking: the complex tone languages are all found either in Subsaharan Africa or in China and Mainland South-East Asia, or New Guinea, with a few representatives in the Americas, particularly Mesoamerica. Eurasia, apart from South-East Asia, is a tone-free area, apart from simple tone systems showing up in a few places (Japan, Scandinavia), as is Australia. For instance in Africa, tone languages are found in all of the major language families: Afro-Asiatic, Niger-Congo, Nilo-Saharan, and Khoisan.

Numeral classifiers (Gil 2011, ch. 55). 400 languages are divided into three types: No numeral classifiers (260), optional (62) and obligatory (78). Obligatory classifiers is a 'Pacific Rim phenomenon' (see Nichols 1992: 198-200, 227-229), mainly found in China and South-East Asia, Japan , Micronesia, Polynesia, Mesoamerica and along the Pacific coast in the Americas. There are also a few representatives in Central- and West Africa, but none in the rest of the world.

Comparative constructions (Stassen 2011, Ch. 121): 167 languages are divided into four types: locational comparatives ('He is old(er) from me'; 78), exceed comparatives ('He is old exceed me'; 33), conjoined comparatives ('He is old, I am young'; 34), and particle comparatives ('He is older than me'; 22). Take the case of the Exceed Comparatives, which "have as their characteristic that the standard NP is constructed as the direct object of a transitive verb with the meaning 'to exceed' or 'to surpass'" (Stassen 2011). They are exclusively found in two areas: Subsaharan Africa and South-East Asia, extending to Polynesia. No examples are found in the rest of the world.[7] The Particle type is basically only found in Europe. According to Stassen, instances of this type in other parts of the world (a few languages in Maritime South-East Asia and the Americas) may be due to

---

[6] Tone is defined as "the use of pitch patterns to distinguish individual words or the grammatical forms of words, such as the singular and plural forms of nouns or different tenses of verbs" (Maddieson 2011).

[7] Newmeyer (1998: 329-330) insists that Classical Greek, Latin and Classical Tibetan "manifest a wide range of comparatives of the 'exceed' type". He takes this as a paradigm example of the problem of relying on secondary sources (descriptive grammars) in typological research. The problem in this case would be that the 'exceed' expressions are not the primary form of comparatives, in these languages, and therefore are not mentioned in standard descriptive grammars.

influence from English or Spanish. The Locative type, on the other hand is distributed more or less evenly across the world.

There are also many features which do not show any clear geographical distribution, for instance articleless NPs (Dryer 2011b), and second position question particles (Dryer 2011c).

These findings are extremely interesting for several reasons. For one thing, they provide interesting evidence of the early history of humans, the spread of populations across the globe and ancient contact between peoples (Nichols 1990, 1992, 1998a,b).[8] But they can also give valuable insights into universal grammar and linguistic variation, as they can show how grammatical properties are more or less susceptible to variation, change, and diffusion (Nichols 1992: 279).

Wichmann and Holman (2009) have carried out an interesting investigation in the relative diachronic stability of the linguistic features in WALS. Their paper first assesses some different methods for modelling stability. They decide in favour of estimating the stability of a feature by "assessing the extent to which phylogenetically related languages are more similar with respect to the feature than are unrelated languages". Applied to the features in WALS, this metric assigns a numerical stability value (in percentage) to each feature. With the range divided four ways (very stable, stable, unstable, and very unstable), and considering the features discussed above, Tone is stable, as are Numeral Classifiers, while Comparative constructions come out as very stable. The word order features discussed earlier (VO vs. OV and NAdj vs. AdjN) also come out as very stable.[9] Among the features which come out as very unstable are, for example, consonant inventories and fixed stress locations (among phonological features), definite and indefinite articles, and features to do with expression of negation (among syntactic features). A notable finding is that the stable features are no less susceptible to diffusion than unstable features; stability means more resistance to change internal to a language, but no more resistance to diffusion.

6.   The problem of limited data

A methodological problem which has always dogged large-scale survey-based research is limited data. These investigations have to rely largely on existing grammars, which can be of highly variable quality, being often based on fieldwork carried out under difficult conditions within a limited amount of time, by researchers who are often without adequate linguistic training. The rules and generalizations are, therefore, often incomplete, vaguely formulated, and insufficiently exemplified.

This problem is less pernicious the more superficial and easily observable the phenomenon is that is under investigation. Consequently typological research is typically engaged with such properties: word stress placement, prepositions or postpositions, question particles in initial, final, or second position, etc. Even so, there is a tendency within typological research to present their results with rather more confidence than they actually warrant. And there is also a tendency within linguistics more generally to treat the results of typological research with more reverence than they

---

[8] Nichols (1992) claims that viewing the distribution of grammatical features as a population-typological or geographical issue "can take us farther back in time than the comparative method can and can see graspable facts and patterns where the comparative-historical method has nothing at all to work with" (p. 280). The comparative method referred to here is the traditional method of tracing genetic relations by comparing cognate morphemes and words. See Longobardi, Gianollo & Guardiano (2008), Longobardi & Guardiano (2009), for extension of the comparative-historical method to syntactic features. One rationale for this extension is, indeed, that it has the potential to take us farther back in history than lexically based comparison.

[9] That order of V and O should be a very stable feature is not obvious from a European perspective, since many of the languages in Europe have recently (in the last 1000 years) undergone change from OV to VO. Apparently this is not such a common phenomenon in a global perspective, though.

deserve (especially when the results happen to support their favoured theory). Obviously, as linguistic theory progresses, with more and more languages being subject to competent, systematic and detailed investigation and description, the more confident we can be that the observations are accurate, and the more abstract the properties can be that are subject to typological research.

For example WALS is quite conservative in the choice of variables that they include, generally keeping to surface- observable phenomena.[10] Even so, for some of the grammatical properties/features that WALS includes, it is questionable whether they lend themselves to the sort of large-scale, surface-based analysis that WALS is based on. I will take an example which I happen to have some knowledge of, having investigated the phenomenon in some detail: The null subject or pro-drop property. I would claim that pro-drop is not a member of the class of easily observable phenomena which can be surveyed and counted on the basis of grammatical descriptions of the usual kind.

WALS includes 'Expression of pronominal subjects' (Dryer 2011d), as one of its 142 features . An impressive total of 674 languages are investigated, and five types are distinguished, plus a mixed type:

1. Obligatory pronouns in subject position (77 languages),
2. Subject affixes on verb (409),
3. Subject clitics on variable hosts (33)
4. Subject pronouns in different position (64)
5. Optional pronoun in subject position (61)
6. Mixed (30)

Type 1 is the typical non-pro-drop type (languages where finite sentences " normally if not obligatorily contain a pronoun in subject position" (Dryer, ibid.). Type 2, by far the commonest, is the agreement pro-drop type, where rich subject agreement substitutes for a subject pronoun or licenses pro-drop (depending on one's theory and analysis; see Holmberg 2005 and below). Type 3 is languages with pronominal clitics, Type 4 has subject pronouns doubling a lexical subject. Type 5 are 'radical pro-drop languages', where subject pronouns can be null without any agreement or other morphological expression of the subject, including Japanese, Thai, etc. Type 6 is a mixed type, which has pro-drop but only in some definable part of the system, as, for example, Finnish, which has pro-drop of 1st and 2nd person pronouns but not 3rd.

The figures probably capture a tendency which is real, but I would not recommend attaching much importance to the exact figures, for reasons discussed below. Type 1 includes not just rigidly pro-drop-free languages, but also languages in which it is "grammatically possible to have simple sentences without anything in subject position, but in which this option is seldom taken in actual usage" (Dryer 2005). This is a wise decision, since probably very few languages completely reject any instances of missing or silent subjects, in finite clauses. For example (spoken) English and the written English of certain clearly definable registers such as diaries does not (see Haegeman 1999), yet there are good reasons to think that the syntax of English null subjects is in important respects different from the syntax of null subjects in, for example, Italian or Arabic. Drawing the line between Type 1, so conceived, and Type 2 is not easy to do in practice, though, on the basis of information typically found in descriptive grammars. I have picked out three among the languages in Type 2 which caught my attention, since I happen to have at least some knowledge of languages in the

---

[10] Furthermore, WALS online is an interactive website, where readers are encouraged to comment on the factual claims. In this way, the typological claims should be gradually improving in accuracy.

relevant families, and because I suspect, on that basis, that they are wrongly classified, or worse, they show that the typology does not make the right distinctions. These are Assamese and Punjabi, two Indo-Aryan languages, and Estonian, a Finno-Ugric language.

Type 2 consists of languages in which "the normal expression of pronominal subjects is by means of affixes on the verb". What does 'normal' mean? For many languages it is easy to determine: The subject pronoun is basically always null except when it is focused/contrasted or expresses a new topic, or perhaps when, for some reason, agreement is unavailable. I would recommend, as a methodological rule of thumb in a case like this, not to rely solely on whatever assertion(s) the author of the grammar makes concerning expression of pronominal subjects, but to check the use of pronominal subjects in examples found throughout the book. If it is a true Type 2 language, examples with pronominal subjects in finite sentences throughout the grammar, especially 1st and 2nd person subjects (as these are typically not dependent on a linguistic antecedent), should have a null subject. If most of them have a spelled-out subject, there is cause to be skeptical. Take the case of Aja and Kresh, two languages included in the list of Type 2 language in WALS, based on Santandrea (1976).[11] The author writes (p. 59) "If the subject is a personal pronoun ... it is regularly omitted, for in most of these languages (including Aja and Kresh – AH), verbs are conjugated, and thus inflexions or tones make up for the omission." The characterisation 'regularly omitted' is quite unambiguous. But in addition, and crucially, this omission is evidenced in sentence after sentence, exemplifying various aspects of the grammar of the languages. The same is found in any grammar of Arabic or Turkish or Spanish, just to mention some more familiar, well-researched Type 2 languages.[12]

Assamese (Asamiya) is also classified as a Type 2 language. In the paper which the classification in WALS is based on (Goswami and Tamuli 2003) I could find no explicit mention relating to null subjects .[13] However, the paper contains a number of examples with pronominal subjects, several of which exemplify varieties of narrow focus on non-subject constituents, hence have a de-focused subject. Only two out of about a dozen examples (p. 482) have a null subject. This is not indicative of a Type 2 language.

As for Punjabi, the work that the classification in WALS is based on is Gill & Gleason (1963). I have access only to the 1969 edition of the book. On the issue of expression of subject pronouns, the book mentions that such pronouns are omitted in connected discourse under topic continuity, and provides an example of a narrative demonstrating this. A search of examples with 1st person singular subjects throughout the book confirms, however, that nearly all of them have an overt pronoun. Bhatia (1993), the most comprehensive syntactic description of Punjabi to date, writes on the expression of subject pronouns: "They are generally dropped if they are traceable either from the verb or from the context." He continues "In non-perfective tenses, the verb agrees with the subject in number, gender, and person; therefore, in such instances, pronouns are often dropped." Bhatia (1993: 222). The qualification 'often' suggests that we may not be dealing with a consistent Type 2 language. But more to the point, examples of finite sentences with pronominal subjects (of which there are many), including those with non-perfective tenses, invariably have a spelled-out subject. It may well be the case that in some of these sentences the subject would be null in natural,

---

[11] I have picked these languages purely for the sake of convenience: My local university library happened to have this book.

[12] See for example Harrell (1962), Kornfilt (1997), XXX.

[13] WALS includes references, with page numbers, for every language-feature ascription. In the case of Assamese-subject pronoun expression the page reference is clearly wrong, though.

connected discourse. Nevertheless, this is indicative of a system different from the one in Aja and Kresh (or Arabic, Turkish, Spanish, etc.).[14]

For Estonian, the WALS reference is de Sivers (1969). On the page referred to (47-48), the personal pronouns of Estonian are listed along with the agreement inflection on the verb, and a set of examples, drawn from a corpus of spoken Estonian. The verb inflection in Estonian is rich enough, with six distinct forms (three persons, two numbers). However, all three (probably) non-emphatic examples of the first person pronoun were overt, and three (probably) non-emphatic examples of the 2sg pronoun were overt, and one null. A consultation with an Estonian colleague (Anne Tamm, p.c.) confirms that Estonian, in particular the spoken language, is by no means a clearcut case of a null-subject language. It is, in this respect, similar to its close relative Finnish. Although written Finnish, conforming to normative pressure, has regular pro-drop of 1st and 2nd person subjects, this is not the case in spoken Finnish except in certain contexts.

There is a perfectly reasonable rationale for the Type 1-Type 2 distinction in Dryer's classification: For a large class of languages the subject agreement inflection on the verb or auxiliary in some sense IS the subject, or at least, can express the subject function on its own: It is (or can be) a pronominal category capable of carrying/expressing the thematic role and case of the subject. These would be the Type 2 languages. This means (a) that there is no null or deleted subject pronoun in the relevant sentences in these languages, and (b) that when there is an overt subject, it is an adjunct, only indirectly linked to the subject role.[15] There are good reasons to think that there are languages like this; see Holmberg (2005), Barbosa (1995, 2009), Alexiadou & Anagnostopoulou (1998). However, in Holmberg (2005) I show that there is at least one language which has rich subject-verb agreement and has overtly subjectless sentences, even quite regularly, but where there is a null or deleted subject, occupying a position in the sentence, namely Finnish. The claim is that this is a more general phenomenon: There is a class of languages, including Finnish, Brazilian Portuguese, and at least some of the Indo-Aryan languages, and probably Estonian, which is called partial null subject languages in Holmberg (2005), Holmberg, Nayudu & Sheehan (2009), Biberauer & al. (2010), distinct from consistent null-subject languages, i.e. the more prototypical Type 2 languages in Dryer's taxonomy. Characteristic of the partial pro-drop languages is that referential subject pro-drop is never obligatory, the way it is in consistent null subject languages when there is a local antecedent and no emphasis on the subject, and generally is syntactically more restricted. Another characteristic of partial null subject languages is that they have (or rather, can have) a null inclusive generic subject pronoun ('null *one*') in active sentences, where the consistent null subject languages always seem to resort to some overt strategy. This classification is based on a detailed investigation of a small number of languages, so far representing only a few language families (Indo-European, Finno-Ugric, Semitic/Afro-Asiatic). The hypothesis proposed is that the class-defining properties are due to a difference in the feature composition of the sentential functional head

---

[14] I am grateful to Dr. Raja Nasim Akhtar (p.c.), a specialist on Punjabi, who also prefers not to classify the language as a null-subject language. Western dialects of Punjabi have a system of pronominal enclitics or suffixes attached to verbs doubling the subject, where "[u]se of pronominal suffixes normally presupposes suppression of the full pronominal forms capable of expressing the same syntactic functions" (Shackle 2005: 673). The examples cited in Shackle's work bear this out. In terms of the WALS taxonomy it should rather be a reason to classify the relevant varieties of Punjabi as Type 3, though.

[15] Dryer (p.c.) rejects the idea that overt subject NPs would be adjuncts in all Type 2 languages. While accepting that it may hold true of "languages with more flexible word order, for languages with more fixed word order, I think the NP is still in some sort of syntactic subject position".

I(NFL), expressed as agreement inflection: In consistent pro-drop languages I has a referential feature , in partial pro-drop languages it does not.[16]

It takes a considerable amount of detailed investigation to establish where such languages stand with respect to expression of subject pronouns. Classification of a language as either a consistent or partial pro-drop language cannot, unfortunately, be done on the basis of token assertions regarding subject pronoun expression in descriptive grammars. Frequency of pro drop may be an indicator: If almost every example sentence with a pronominal subject in the grammar (particularly 1st and 2nd person, which are not to the same degree context-dependent as 3rd person) has a null subject, then we are probably dealing with a consistent pro-drop language, a canonical Type 2 language in Dryer's taxonomy.[17] But what if half of them do? Given that the null-subject option exists, as it does in most languages, there are a variety of factors, including sociolinguistic factors, which may influence how this option is employed in spoken or written language, so frequency is at best a weak indicator of class-membership.

The partial pro-drop languages should also be kept distinct from languages like English, which allow null subjects in certain syntactic contexts, and employs them, even quite frequently, in ordinary discourse, subject to regional, social and stylistic variation. Characteristic of English and the other Germanic language is that the null subjects do not occur in embedded contexts (Haegeman 1999, Haegeman & Ihsane 1999). This is the sort of observation which may very well escape a researcher writing the first grammatical description of a language, and often will not be gleanable from the typically scarce example sentences in the grammar. This will hardly be the case in grammars written today, but is perfectly possible in the case of grammars written , say, fifty years ago.

Two conclusions follow from the discussion above: The first point is that the taxonomy in WALS makes too coarse a distinction between Type 1 (languages where sentences with pronominal subjects "normally if not obligatorily" have an overt pronoun in subject position) and Type 2 languages. The second is the methodological point that 'expression of subject pronouns' is too abstract a phenomenon to be subject to a broad-based typological investigation, at the present time, with the descriptions we presently have at our disposal of most languages.

The online interactive database *Syntactic Structure of the World's Languages* (SSWL) implements a different approach from WALS intended to avoid the problem of limited data. It is wholly based on information supplied by language experts directly to the database. Second, the information about syntactic features is strictly coded, according to a strict protocol, as 'yes' (the language has the property) or 'no' (it does not have the property). This means avoiding, as far as possible, any indeterminacy and ambiguity in the use of grammatical terminology, and also means that the database provides explicit negative evidence. The drawback is that the database is wholly dependent on voluntary contributions from linguistic experts, and the development of the database has therefore been rather slow.

---

[16] See also Phimsawat (in prep.) who shows that Thai, along with several other languages of Type 5 in WALS, that is 'radical pro-drop languages' with extensive pro-drop but no agreement, share with partial pro-drop languages the property that the (inclusive) generic pronoun is null. This is predicted by the theory in Holmberg (2001, 2010): They have no nominal features in I, hence no referential feature.

[17] Finnish happens to be an exception, though: Due to normative pressure in the written language, 1st and 2nd person null subjects are massively over-represented in examples in descriptive grammars (e.g. Sulkala & Karjalainen 1992)

7.   Some recent trends[18]

In many ways the research program of Greenbergian linguistic typology has been a success story. As a result of it we now have a vast amount of knowledge about grammatical properties of a large part of the existing languages of the world, including detailed knowledge about common properties and existing variation within most observable subsystems of grammar, phonological, morphological, and syntactic. The extent of large-scale areal spread of linguistic features has been discovered. On the other hand the research program can be said to have failed to achieve one of its goals, that of uncovering universally valid and theoretically interesting generalizations on the basis of systematic comparison of large numbers of languages. As a consequence, an influential school of thought within modern typological research has all but given up the quest for universals along Greenberg's lines, settling instead for mere description of patterns of variation.

> Large datasets almost invariably reveal exceptions to universals , and this, together with a substantial increase of newly described languages assisted by prominent conceptual argumentation (e.g. Dryer 1998, Croft 2002: Chapter 8) has practically done away with notions of absolute universals and impossibilities. Modern studies of typological distributions involve statistical methods, from association tests to multivariate scaling methods ... The general assumption is that if there are large-scale connections between linguistic structures, or between linguistic structures and geography, they consist in probabilistic (and therefore exception-ridden) correlations between independently measured variables. (Bickel 2007: 245)

Some recent, radical representatives of this view which have received a certain amount of attention are Evans & Levinson (2009) and Dunn & al. (2011). These works claim to demonstrate the absence of formal or function-based universals of any kind. Evans & Levinson work their way through a set of grammatical concepts and generalizations which, they say, have at one time or other been claimed to have universal validity, purporting to show that there are counterexamples to all of them, and that they are therefore disconfirmed as theoretical hypotheses. Their conclusion is that there is no evidence of a biological basis for linguistic properties. Dunn & al. (2011) make essentially the same point, purporting to show specifically that there is no evidence for the Greenbergian word order universals, not even the reduced set in Dryer (1992). Instead, they argue, the proposed correlations of the sort that researchers in the Greenbergian tradition typically have assumed follow family lines. The conclusion is that variation is unpredictable and unconstrained, except for genealogical/historical constraints (related languages have more features in common than unrelated languages).[19]

In this context, Baker (2010) makes the following important point. The Chomskyan view of UG is that it is a property of human cognition which makes possible acquisition of one or more languages, each of them a staggeringly complex system of lexical items, categories, rules, and principles, in the space of a few years, on the basis of ordinary linguistic experience with its well known limitations (absence of negative data, highly variable exposure to data, and absence of

---

[18] The content in this section is heavily based on Baker (2010).
[19] See Baker (to appear) for a discussion of Dunn & al. (2011) demonstrating that their overall conclusions do not follow from the statistical facts they discover. See the comments on Evans & Levinson's (2009) target paper, many of which show that they are massively overstating their case.

instruction). UG is 'the initial state' which, when provided with primary data eventually yields the final state, the adult grammar and lexicon, which makes possible use of language for communication and verbalized thought. Variation among languages is due to variation in the primary data, but the variation is restricted by various factors, one of which is properties of the genetically determined 'initial state', UG. Chomsky's more recent position is that the initial state is less a matter of UG in the sense of a cognitive faculty dedicated to language, and more a matter of general properties of computational/combinatorial systems, but nevertheless, there is a biologically given, mind-internal faculty which makes possible acquisition of a language, given sufficient primary data, as found in any normal human society. But since there is variation in the data (basically as much as is allowed by UG and other constraining factors), the outcome is different languages and dialects.

As discussed, typological research in the Greenbergian tradition is, for practical reasons, largely restricted to surface-observable properties, such as word order and overt morphological properties. These properties are observable to the descriptive grammarian, so they are also observable to the language learner in the data he/she encounters. That means that they can, on that account, be learnt on the basis of experience. This, in turn, means that they are among the linguistic facts that we can expect to vary across languages. The properties that we expect *not* to vary are all those abstract properties which are *not* directly observable in the primary data. That is to say, the properties, rules, categories, etc. that are best amenable to Greenbergian typological investigation are precisely those where we expect *not* to find absolute, exceptionless generalizations, but at most tendencies of varying strength.

Correspondingly, from a Chomskyan perspective, we expect to find absolute generalizations regarding properties that are more abstract, not directly observable in the data. Baker (2010) mentions the existence of VP as an example of a (potential) universal, which is not directly observable in the primary data, since most languages have a variety of constructions where the verb and the object are separated (in VSO languages the subject systematically intervenes between the verb and its object). Another case that he mentions is island constraints on movement, such as the complex NP constraint and the coordinate structure constraint. Island constraints can basically only be detected by negative examples, and these do not occur in ordinary linguistic experience, yet there are good reasons to think that they are universal, presumably reflecting the limits of human linguistic computing capacity, even though the precise formulation of the various islands is still open to debate. Even more fundamental and more abstract properties that are, in all likelihood, universal are hierarchic syntactic structure (possibly strictly binary branching), and the role of c-command as a regulator of scope of operators and modifiers as well as movement, binding, and agreement relations.[20] There may be some variation in the opacity of certain categorial boundaries, and the precise formulation of c-command is still open to debate and refinement, but there is very good reason to think that the relation plays essentially the same crucial role in every language.

Consider the following quotation from Haspelmath (2008):

> But there does seem to be a widespread sense in the field of (non-generative) typology that cross-domain correlations do not exist and should not really be expected. After the initial success of word order typology, there have been many attempts to link word order (especially VO/OV) to other aspects of language structure.... But such attempts have either

---

[20] Needless to say, these notions are still subject to debate in the literature. See for example Nunes (2004) on c-command and movement, and Culicover & Jackendoff, (2005) on the role of c-command generally.

failed completely, or have produced only weak correlations that are hard to distinguish from areal effects. (Haspelmath 2008: 95)[21]

There is a reason for the failure to find exceptionless correlations based on VO/OV order: We expect to find such correlations at the level of syntactic structure, including syntactic features and operations. But surface word order is only a weak predictor of syntactic structure. SVO order can be derived in a variety of ways (for example by V-movement in a language which otherwise has SOV order , as in German and Vata; see Koopman (1994), as can SOV order. Even for languages which are rigidly either SOV or SVO there is no a priori reason to think that they necessarily have identical sentential syntactic structure; see Holmberg (1998).

V-initial order can also be derived in a variety of ways. There are very good reasons to think that there are at least two entirely different types of V-initial order among the languages of the world (see Carnie, Dooley & Harley 2005). One type is exemplified by Celtic and Mediterranean VS order (Berber, Arabic, Spanish, Greek). In these languages VS order is derived by V-movement. A different type is exemplified by Niuean and (very likely) other Oceanic (and perhaps more generally Austronesian) VS languages. As shown by Massam (2001, 2005), Niuean not just V-initial, but more generally predicate-initial. VS order is not derived by V-movement, but by VP-movement, which yields an entirely different syntactic structure, but often the same word order as V-movement.[22] From a syntactic point of view, it is not obvious that the two VS types would systematically share *any* properties that they would not share with other languages, apart from (sometimes) identical surface order, while we do expect V-initial languages which derive VS order by V-movement to (possibly) share certain properties, namely properties that are linked to V-movement.

A point made by Dryer (1998b) and also Newmeyer (1998) is that we can never establish absolute universals on the basis of just cross-linguistic comparison, simply because now existing languages constitute only a fraction of all possible languages: Countless languages have disappeared without a trace, and countless language have not yet appeared. What we are expecting from cross-linguistic comparison is to give us clues about possible linguistic universals. Establishing whether they are truly universal, in the sense of being necessary components of human cognition, requires other kinds of data, for example experimental data showing that a particular structure or rule is unlearnable under natural conditions.

8.   Comparing categories across languages

It is crucial for any comparative cross-linguistic research to know that we are comparing like with like. Does the subject in a given English sentence have the same structural import as the subject in the roughly corresponding Arabic sentence? Is the adposition meaning 'on' in English the same category as the case-suffix meaning 'on' in Finnish?

---

[21] Haspelmath (2008) does not claim that there are no universals. His argument concerns what he calls cross-domain universals, by which he means correlations between phonological and syntactic properties, or between VP-properties and NP-properties, or order in the VP and type of negation, etc. etc. Universals that do exists are what he calls intra-domain universals. They typically have the form of universal hierarchies or scales, where languages can vary with regard to the placement of a feature. Keenan & Comrie's (1977) accessibility hierarchy for relativization is one example: The hierarchy is: Subject > Direct Object > Oblique > Possessor. If a language can relativize on any position on this hierarchy it can also relativize on every higher position (Haspelmath 2008: 96).

[22] See Chung (2005), Holmer (2005), Otsuka (2005) for discussion of VP-raising as a source of V-initial order in a variety of Austronesian languages.

Ever since Greenberg it has been taken for granted, or at least, it has been assumed as a heuristic in typological research that formal, syntactic and phonological categories can be compared across languages; this is basis for most of the research behind WALS, for example. Greenberg (1963: 74) concedes that he is basically employing semantic criteria when identifying grammatical relations and word classes. He notes that there are "very probably formal similarities which permit us to equate such phenomena in different languages" (ibid.), but refrains from entering any such discussion, adding "In fact there was never any real doubt in the languages treated about such matters"(ibid.).

In more recent years the issue has, again, come under scrutiny, and influential typologists such as Matthew Dryer, William Croft, and Martin Haspelmath have argued against the existence of cross-linguistic formal/structural categories. There are obviously similarities among grammatical items and categories across languages, but not identity. In a sense, this is almost a return to the position of the American structuralists, as most famously articulated in Boas 1911), who claimed that each language was a system onto itself, describable only on its own terms (see Croft 2001: 34). It is not quite a return to this position, since the modern functional typologists obviously do not deny the possibility and usefulness of comparative research. The claim is that the comparison can be, and should be, done on the basis of 'comparative concepts' (Haspelmath's 2008 term), that is categories which are drawn from traditional and/or formal grammar , but which are only loosely defined, typically on the basis of meaning or function, in a pragmatic fashion, the definitions being only just formal enough to permit stating broad cross-linguistic generalisations.

 For example 'adjective' as a comparative concept is defined as "a lexeme that is primarily used to attribute a property to a nominal referent" (Haspelmath). Crucially, the definition does not specify the categorial features of the word (verb, noun, or neither). As such, cross-linguistic generalisations can be stated over this comparative concept, including Greenberg's word order universals (see above) and Cinque's (2008) hierarchic universals, which hold true whether the words in question are verbal , nominal or neither; see Dryer (forthcoming) for arguments that this is empirically correct. Thus while noun-modifying items in individual languages have particular categorial features, they are also realisations of a broader, language-neutral 'comparative concept'.

See Newmeyer (2007) for a critique of the claim that typology can, and should, do without formal categories. Newmeyer makes essentially two points: First, that typological research has always relied on formal categories, with considerable success, and still does so, as the categories now called 'comparative concepts' are still in part formally defined, even when the definitions are eclectic and purposely ad hoc. Second, replacing formal syntactic and morphological criteria by semantic criteria, as advocated by Dryer (2007) and Haspelmath (2008), has its own problems. The argument is that semantic categories are (a) universal and (b) not subject to theoretical controversy the way syntactic categories are. At least the last claim is wrong, Newmeyer points out, as semantic theory is every bit as controversial as syntactic theory.[23]

The obvious optimal solution would seem to be that development of formal syntactic theory proceeds hand in hand with typological investigation. Formal syntactic theory needs reliable comparative data to test its hypotheses, but this will typically require finer formal distinctions to be made than is used in WALS, for example. And if the ambition, on the typologists' side, is to get beyond the present state of understanding of linguistic variation, this may well require finer formal

---

[23] Greenberg's (1963) criteria are in fact only in part semantic. The word order universals would lose much of their generality if the notions 'subject' or 'genitive', for example, were defined just on semantic grounds; see Newmeyer (2007).

distinction, too. For example, proceeding from the typology reported in Dryer (2011d) concerning expression of subject pronouns to a finer classification of Type 2 (along the lines suggested above), would seem to require a more detailed analysis of the subject-verb agreement element, presumably in terms of ϕ-features, arguably including features such as [±referential] or [±definite]. This is not to deny that certain cross-linguistic generalizations may actually be most appropriately stated in terms of semantically defined categories (as argued in Dryer, forthcoming).

9. The Middle Way: selective global comparison

How big does a sample of languages need to be, to serve as a viable testing ground for hypotheses about the nature of language? Given the limited scope of descriptive grammars and the almost inescapable unreliability of the data and the generalisations in them, it would be an advantage if careful sampling could substitute for large numbers. Let us say, instead of 100 or 1000 languages selected for the comparison of a particular feature, there would be 10 carefully selected languages. There are two obvious advantages: First, the languages in the sample could then be subjected to more detailed scrutiny, to avoid descriptive gaps and false generalizations, and to make sure that the descriptive framework used to describe the languages, and the theoretical assumptions underlying it, are identical, or at least compatible (on this problem, see Newmeyer 1998: 337ff.). Second, the investigation would not need to be restricted to 'surface observable' properties; ideally negative data would be available as well. But importantly, selection of the ten languages would be such that they represent the different (major) families and different areas of the world, thus avoiding any obvious genetic or geographical bias. This method is advocated by Baker & McCloskey (2007), under the name of the Middle Way. They see it as way to combine the virtues of linguistic typology with generative linguistic research, hopefully with results which will be seen as interesting and relevant to both research programs.

An additional sampling criterion would be that the restricted sample should include a variety of types expected to show variation with respect to the feature investigated. Assume that the investigation is about generic pronouns, a topic which I happen to be interested in. The sample should include one or more languages of Type 1 in Dryer (2011d) (non-pro-drop languages) as well as Type 2 (pro-drop languages), and among the Type 2 languages, consistent as well as partial ones (see section 6 above). There should be pro-drop languages without agreement (Dryer's Type 5). There should be languages with an overt dedicated generic pronoun, like English *one*, and languages where some other overt indefinite pronoun ('anyone') is employed to convey generic meaning. It should include one or more languages with a null generic pronoun with no special morphological conditions such as a particular impersonal verb-inflection, and preferably one where a null generic pronoun is licensed by impersonal inflection. Unless the investigation is restricted to generic *subject* pronouns, the principal variation among object generic pronouns should also be covered (null vs. pronounced, dedicated pronoun or not, object agreement or not). These typological characteristics will overlap in various ways. The sample is based in part on prior knowledge of what factors make a difference, in part on educated guesswork. The expectation is that there will be patterns in the variation discernible even in a small but carefully selected sample, which can be probed further in the languages under investigation, possible because the number of languages is small, and/or can be tested on a larger sample.

This method, combining a genetically and regionally balanced global mini-sample with a directed search for representatives of a variety of types, presupposes some knowledge of the distribution and the variation of the feature that we are interested in among the languages of the

world, as well as a certain amount of knowledge of the formal properties of the phenomenon investigated. That is to say, it presupposes investigation of the large-scale comparative type, as a first step, as well as a preliminary investigation of the formal properties of the phenomenon, preferably in more than one or two languages, to have a preliminary idea of what properties you want to include in a broader comparative investigation. This is, then, a kind of investigation which can be done successfully, at this stage, by a combination of the methods of linguistic typology and generative linguistics.

In fact, something like this method, combining a broad survey with a more directed investigation of a smaller sample, has been employed in typological research. Haspelmath (1997) is a good example. In his investigation of indefinite pronouns he employs two samples. One consists of 40 languages, selected because there are grammatical descriptions of these languages which are detailed enough and reliable enough for the purposes of his investigation, and in addition native speakers can be contacted. This sample is heavily skewed towards European languages. The other sample contains 100 languages, and is sampled so as to be representative of the world's languages (Haspelmath 1997: 16-17). The questions that are asked from these two samples are correspondingly different: The big sample can only supply information about a few superficial properties, while the small sample can provide information on issues of a more abstract nature. Haspelmath points out that indefinite pronouns appear to be a diachronically unstable phenomenon, exhibiting considerable variation even among closely related languages, and for this reason, too, use of a genetically and geographically biased sample is not as detrimental as might otherwise be the case (ibid.: 16).[24] In fact, the variability of a feature on the micro-level (between closely related languages and dialects) is the sort of information which can easily go unnoticed in an investigation based on a globally representative sample – and which may well turn out to be important for the proper understanding of the phenomenon. This indicates that Middle Way type of investigations based on globally representative mini-samples may need to be combined with micro-level investigation to minimize the risk of missing important generalizations.

10.   In conclusion

Discussing the difference between research on language universals and linguistic typology, Comrie (1989) writes "[T]he only difference [is] that language universals research is concerned primarily with limits on [the variation within human language -- AH], whereas typological research is concerned more directly with possible variation". And he continues "However, neither conceptually not methodologically is it possible to isolate the one study from the other." (Comrie 1989: 34). In this vein, I have stressed the complementary nature of research in typological linguistics and generative, formal linguistics, in this paper. The great strength of the typological programme is in the method of investigation, the large-scale comparison, these days supported by increasingly sophisticated methods of processing data. The successes of the programme in this regard are undisputable. There are also some rather obvious problems pertaining to large-scale comparative investigations, in part stemming from the fact that few languages have so far been described in sufficient detail to allow large-scale comparison of more abstract grammatical properties. The functionalist orientation which is characteristic of the linguistic typology stems at least in part from

---

[24] On the other hand indefinite pronoun systems show a clear continent-size areal pattern, with interrogative-based indefinite pronouns in the languages of Eurasia, America, and Australia, and generic-noun-based indefinites in Africa and Oceania, which appears to contradict the claim that they are diachronically unstable (Haspelmath 1997: 241).

conditions on the method. If the properties compared are necessarily surface-observable, they will thereby be subject to some degree of unpredictable variation. Generalizations will therefore be exception-ridden, and thus often easier to explain in functional than formal terms.

What the right explanation is, in any given case, is an empirical issue, though. I happily agree with Hawkins (to appear), Haspelmath (2008), and Newmeyer (2005) that it is in the interest of everyone to distinguish between functionally and formally motivated properties of grammar. If a phenomenon has a credible functional explanation, then we can move on from there and look at the next phenomenon, without prejudice.

There is no principled reason why generative, formally oriented linguistics should not engage in broad-based comparison, following the methodology of modern typological research. There are some good examples of this practice. Cinque (1999) on adverbs and Cinque (2009) on adjectives, and Julien (2002) on tense and aspect marking are three well known examples. Each of them is based on surveys of several hundred languages, with a view to constructing a universally valid theory of the functional sequence in the sentence (Cinque 1999, Julien 2002) and the noun phrase (Cinque 2009), with consequences for formal hypotheses such as the Mirror Principle (Baker 1989) and the status of the notion 'word' in syntax. Giuseppe Longobardi's project researching genealogical relations based on mass comparison of syntactic properties in large sets of languages could also be mentioned here (Longobardi, Gianollo & Guardiano 2008, Longobardi & Guardiano 2009). Unfortunately these works remain rather isolated examples. There are some signs of increased, serious interest in typology within generative grammar, though, such as the ReCoS project at Cambridge (Roberts 2011). The SSWL database, created by Chris Collins and Richard Kayne, is also a step in that direction.

References

Alexiadou, A. and E. Anagnostopoulou 1998. Parametrizing AGR: word order, V-movement and EPP checking. *Natural Language and Linguistic Theory* 16: 491-539.

Baker, Mark. 1985. The Mirror Principle and morphosyntactic explanation. *Linguistic Inquiry* 16: 373-416.

Baker, Mark. 2010. Formal Generative Typology. In B. Heine and H. Narrog (eds.) *The Oxford Handbook of Linguistic Analysis*, Oxford University Press, Oxford, pp. 285-312.

Baker, Mark. To appear. The interplay between Universal Grammar, Universals and Lineage Specificity: Some reflections raised by Dunn, Greenhill, Levinson and Gray 2011. To appear in *Linguistic Typology*.

Baker, Mark and James McCloskey. 2007. On the Relationship of Typology to Theoretical Syntax. *Linguistic Typology* 11:273-284.

Barbosa, P. 1995. *Null subjects.* Ph.D. dissertation: MIT.

Barbosa, P. 2009. Two kinds of subject *pro*. *Studia Linguistica* 63. Special issue: Partial pro-drop (ed. by A. Holmberg): 2-58.

Bhatia 1993. Punjabi

Biberauer , T., A. Holmberg, I. Roberts, M. Sheehan. 2010. *Parametric variation. Null subjects in minimalist theory.* Cambridge University Press.

Bickel , B. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11: 239-251.

Boas 1911.

Bobaljik, J. (2002b) Realizing Germanic inflection: Why morphology does not drive syntax. *Journal of Comparative Germanic Linguistics*, special issue on Lexical Integrity [eds. Ackema, P. & Neeleman, A.], 6, pp. 129-167.

Carnie, A., S. Dooley and H. Harley. 2005. *Verb first*. Philadelphia: John Benjamins.

Chomsky, N. 1981. *Lectures on Government and Binding.* Dordrecht: Foris.

Chomsky, N. 1982. *Some concepts and consequences of the theory of Government and Binding.* Cambridge, MA: MIT Press.

Chomsky, N. 1986. *Knowledge of language: its nature, origin, and use* Praeger: New York.

Chomsky, N. 2005. Three factors in language design. *Linguistic Inquiry* 36: 1-22.

Chung, S. 2005. What fronts? On the VP raising account of Verb-Initial Order. In A. Carnie, H. Harley and S. Dooley (eds.). *On the syntax of verb-initial languages.* Amsterdam: Benjamins, 9-29.

Cinque, G. 1999. *Adverbs and functional heads. A crosslinguistic perspective.* New York: Oxford University Press.

Cinque, G. 2009. *The syntax of adjectives. A comparative study.* Cambridge MA: MIT Press.

Comrie, B. 1989. *Language universals and linguistic typology*. University of Chicago Press.

Croft 1995. Autonomy and functionalist linguistics. *Language* 71: 490-532.

Croft, W. 1999. The Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species, Cambridge University Press, 2nd rev. edition .

Culicover, P.W. and Jackendoff, R. 2005). *Simpler Syntax*. Oxford University Press.

Dryer, M. S. 1988. Object-verb order and adjective-noun order: dispelling a myth. *Lingua, 74*, 185-217.

Dryer, M.S. 1989. Large linguistic areas and language sampling. *Studies in Language 13*, 257-292.

Dryer, M. S. 1992. The Greenbergian Word Order Correlations. *Language, 68*, 81-138.

Dryer, M.S. 1997. Are grammatical relations universal? In: *Essays on language function and language type: Dedicated to T. Givón*, ed. by J. Bybee, J. Haiman, & S.A. Thompson, 115-143. Amsterdam: Benjamins.

Dryer, M.S. 1998a. Aspects of word order in the languages of Europe. In Siewierska, A. (ed.) *Constituent order in the languages of Europe*, 283-320. Berlin: Walter de Gruyter.

Dryer, M. 1998b. Why statistical universals are better than absolute universals. *Chicago Linguistic Society 33: The Panels*, pp. 123-145.

Dryer, M.S. 2007 The order of demonstrative, adjective, numeral, and noun: An alternative to Cinque. Talk presented at *Workshop on theoretical approaches to disharmonic word orders*, Newcastle University.

Dryer, M. 2011a. Genealogical Language List. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Dryer, M. 2011b. Definite articles. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Dryer, M. 2011c. Position of polar question particles. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Dryer, M. 2011d. Expression of pronominal subjects. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Dunn, M. , S.J. Greenhill, S. C. Levinson & R.D. Gray. 2011 Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473: 79-82.

Evans, N. & S.C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioural and Brain Sciences* 32: 429-492.

Gil, 2011. Numeral classifiers. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Gill, H.S. and H.A. Gleason (1963/1969). *A reference grammar of Panjabi* [revised edition]. Department of Linguistics, Punjabi University.

Goswami, G. C. and J. Tamuli 2003. Asamiya.In G. Cardona and D. Jain (eds.) *The Indo-Aryan languages*. London: Routledge.

Greenberg, J. 1963/2007. Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (ed.). *Universals of Language*. Cambridge, MA: MIT Press, 58-90. Reprinted in I. Roberts (ed.). 2007. *Comparative grammar*, Volume 1, 41-74.

Greenberg, Joseph (1974) *Language typology: a historical and analytic overview*. The Hague: Mouton.

Greenberg, J., Charles A. Ferguson & Edith A. Moravcsik (eds.). 1978. *Universals of human language*. Stanford, CA : Stanford University Press, 1978.

Haegeman, L., (1999) 'Adult null subjects in non pro-drop languages', in Friedemann, M-A. &
L. Rizzi eds., *The Acquisition of Syntax: studies in comparative developmental
linguistics*, London: Longman.

Haegeman, L. & Ihsane. 1999. Subject ellipsis in embedded clauses in English. *English
Language and Linguistics* 3: 117-145.

Harrell, Richard S. 1962. *A short reference grammar of Moroccan Arabic*. Washington DC:
Georgetown University Press.

Haspelmath, M. 1997. *Indefinite pronouns*. Oxford: Clarendon Press.

Haspelmath, M. 2008. Parametric versus functional explanations of syntactic universals. In
Biberauer, T. (ed.) *The limits of syntactic variation*, 75-108. Amsterdam: John
Benjamins.

Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2005. *The world atlas of language
sructures*. Oxford: Blackwell.

Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011. *The world atlas of language
structures*. Munich: Max Planck Digital Library. Available online at http://wals.info/

Hauser, M., N. Chomsky and T. Fitch. 2002. The faculty of language: what is it, who has it, and how
did it evolve? *Science* 298: 1569-79.

Hawkins, John A. 1979. Implicational universals as predictors of language change. *Language*
55: 618-648.

Hawkins, John A. 1983. *Word order universals*. New Yourk: Academic Press.

Hawkins, J. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University
Press.

Hawkins, J. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John A. To appear. Formal and functional approaches to disharmonic word order. To
appear in Biberauer, T & M. Sheehan (eds.) *Theoretical Approaches to Disharmonic Word
Orders*. Oxford University Press.

Holmberg, A. 1998. Word order variation in some European SVO languages: a parametric approach.
In Siewierska, A. (ed.) *Constituent order in the languages of Europe*. 553-598. Berlin:
Walter de Gruyter.

Holmberg, A. 2005. Is there a little *pro*? Evidence from Finnish. *Linguistic Inquiry* 36: 533-64.

Holmberg, A., A. Nayudu & M. Sheehan. 2009. Three partial null-subject languages: a comparison of
Brazilian Portuguese, Finnish, and Marathi *Studia Linguistica* 63. Special issue: Partial pro-drop
(ed. by A. Holmberg).

Holmer, A. 2005. Seediq: Antisymmetry and final particles in a Formosan VOS language. In A. Carnie, H. Harley and S. Dooley (eds.). *On the syntax of verb-initial language,* 175-202. Amsterdam: John Benjamins.

Julien, M. 2002. *Syntactic heads and word formation*. New York: Oxford University Press.

Keenan, E. & B. Comrie 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8: 63-99.

Koopman, H. 1984. *The syntax of verbs*. Dordrecht: Focus.

Kornfilt, J. 1997. *Turkish*. London: Routledge.

Lehmann, W. 1973. A structural principle of language and its implications. *Language* 49: 47-66.

Lehmann, W. (ed.) 1978. *Syntactic typology: studies in the phenomenology of language*. Austin: University of Texas Press.

Longobardi, G., C. Gianollo, C. Guardiano. 2008.Three fundamental issues in parametric linguistics, (in collaboration with C. Gianollo and C. Guardiano), in T. Biberauer (ed.) *The Limits of Syntactic Variation,* , 109-142. John Benjamins, Philadelphia/Amsterdam, 2008.

Longobardi, G. and C. Guardiano. 2009. Evidence for Syntax as a Signal of Historical Relatedness, (in collaboration with C. Guardiano), in *Lingua* 119, 11, special issue: The Forests behind the Trees, ed. by John Nerbonne, 1679-1706.

Maddison 2011. Tone. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Massam, D. (2001) On predication and the status of subjects in Niuean. In W. Davies and S. Dubinsky (eds.). *Objects and other subjects: grammatical functions, functional categories, and configurationality.* Dordrecht: Kluwer, 225-246.

Massam, D. (2005) Lexical categories, lack of inflection, and predicate fronting in Niuean. In: Carnie, A., Dooley, S. & Harley, H. (eds). *Verb First. On the syntax of verb-initial language* (pp. 227-242). Amsterdam: Benjamins.

Newmeyer, F. 1998. *Language form and language function.* Cambridge MA: MIT Press.

Newmeyer, F. 2005. *Possible and probable languages. A generative perspective on linguistic typology.* Oxford: Oxford University Press.

Newmeyer, F. 2007. Linguistic typology requires crosslinguistic formal categeories. *Linguistic Typology* 11: 133-157.

Nichols, J. 1990. Linguistic diversity and the first settlement of the New World. *Language* 66:3.475-521.

Nichols, J. 1992. *Linguistic diversity in space and time.* Chicago: University of Chicago Press.

Nichols, J. 1998a. The Eurasian spread zone and the Indo-European dispersal. In R. M. Blench et al., eds., *Archaeology and Language II: Correlating Archaeological and Linguistic Hypotheses*, 220-66. Routledge.

Nichols, J. 1998b. Amerind personal pronouns: A reply to Campbell. (With David A. Peterson.) *Language* 74:3.605-15.

Nunes, Jairo. 2004. *Linearization of Chains and Sideward Movement* . MIT Press.

Otsuka. 2005. Two derivations of VSO: A comparative study of Niuean and Tongan. In A. Carnie, H. Harley and S. Dooley (eds.). *On the syntax of verb-initial language,* 65-90. Amsterdam: John Benjamins.

Phimsawat, O. Forthcoming. *The syntax of pro-drop in Thai.* PhD, Newcastle University.

Piattelli-Palmarini, Massimo & Juan Uriagereka. 2008. Still a bridge too far? Biolinguistic questions for grounding language on brains. *Physics of Life Reviews* 5: 207-224. Available online at www.sciencedirect.com.

Rijkhoff, J. and D. Bakker. 1998. Language sampling. *Linguistic Typology* 2: 263-314.

Santandrea, S. 1976. *The Kresh group, Aja and Baka languages (Sudan): a linguistic contribution*. Napoli: Instituto universitario orientale.

Siewierska, A. & D. Bakker (1996) 'The distribution of subject and object agreement and word order type. *Studies in Language* 20: 115-161.

de Sivers, F. 1969. *Analyse grammaticale de l'estonien parlé*. Clermont-Ferrand: G. de Bussac.

Stassen, 2011. Comparative constructions. In Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2011.

Stowell, T. 1981. *The origins of phrase structure*. Ph.D. dissertation: MIT.

Sulkala & Karjalainen 1992. *Finnish*. London: Routledge.

Travis, L. 1984. *Parameters and effects of word order variation*. Ph.D. dissertation: MIT.

Venneman, Theo. 1974. Topics, subjects, and word order: From SXV to SVX via TVX. In John Anderson & Charles Jones (eds.) *Historical linguistics*, 339-376. Amsterdam: North Holland.

Venneman, Theo. 1984. Typology, universals, and change of language. In J. Fisiak (ed.) *Historical syntax.* Trends in Linguistics: Studies and Monographs 23, 593-612. Berlin: Mouton.

Wichmann, Søren and Eric W. Holman. 2009. *Assessing Temporal Stability for Linguistic Typological Features.* München: LINCOM Europa. [PDF]