Economy in formulating typological generalizations

VLADIMIR PERICLIEV

Abstract

Typically, a linguistic typology defines all logically possible types and states which of these types are actually attested and which are not. The task then is to describe such a typology, preferably in the most economic way. In this paper, a descriptive principle is justified to the effect that for any typology with at least one unattested type there exists a minimal description, consisting of a conjunction of non-statistical (implicational) universals, defining all and only the attested types. A method is proposed that finds the minimal description(s) of a typology, and a computer program is sketched that executes the method, illustrating it on the typologies in Greenberg’s Appendix II (1966) and Hawkins’ Expanded Sample. Hawkins (1983) has noticed that Greenberg’s universals do not describe all and only the attested types in Appendix II, but our analysis shows that Hawkins himself has not been fully successful in describing the typology in his Expanded Sample either.

Keywords: computer-aided universals formulation, implicational universals, logic of typology, methodology, word order

1. Introduction

A linguistic typology states all logically possible types, and is typically accompanied by empirical facts as to which of these logically possible types are actually attested and which are not attested (relative to a sample). Given such a typology, linguists face the task of describing it by formulating statements that capture the distributional pattern shown by the typology. The typology’s description would normally be required to be as simple as possible, in accordance with the meta-scientific principle of simplicity (simplex sigillum veri), which is widely endorsed in linguistics. The question arises of how we can find the simplest description(s) of a typology.
The basic goal of this paper is twofold. First, we postulate a descriptive typological principle to the effect that for any typology with at least one unattested type there exists a minimal description, consisting of a conjunction of non-statistical (implicational) universals defining all and only the attested types. No non-statistical generalizations are possible for typologies in which all types are attested. And, secondly, we propose a method that finds the simplest description(s) of a typology, alongside with a computer program that executes the method.

Our discussion is organized as follows. Section 2 reviews the analyses of the word order typologies in Greenberg’s Appendix II (Greenberg 1966) and Hawkins’ Expanded Sample (Hawkins 1983) as familiar illustrations of the kind of task we are discussing. Neither author has proposed a set of universals defining all and only the attested types in his data. In Section 3, using some results from propositional logic, we postulate a principle stating the existence of a minimal account of typologies. This principle allows us to make the a priori judgment that both Greenberg’s and Hawkins’ analyses are, from a descriptive point of view, not fully adequate. Section 4 outlines a method for finding minimal account(s) of a typology in terms of implicational universals, which is illustrated on Hawkins’ typology; the Expanded Sample turns out to have a number of alternative accounts, three of which are simplest in that they consist of the smallest number of universals. In this section, we also list the possible (simplest and non-simplest) accounts of Appendix II. Section 5 is a brief sketch of MINTYP, a computer program that executes the method, which is computationally complex and hence requires automation. And, finally, Section 6 summarizes our contributions.

2. Greenberg’s Appendix II and Hawkins’ Expanded Sample

In his seminal paper on the order of meaningful elements, Greenberg (1966) proposes a typology of the world’s languages in terms of their ordering properties, which he states as Appendix II in his article. (For a good review of the state of the art in word order typology, cf. Dryer 1995.) This typology uses four dimensions (= properties of languages), viz. verb–subject–object order, adposition order (preposition or postposition), adjective–noun order, and genitive–noun order. Greenberg notes that only three out of the six logically admissible types of verb–subject–object order are attested, viz., SVO, SOV, and VSO, while the others are non-existent or extremely rare. Then, assuming binary
Table 1. The Typology in Greenberg’s Appendix II

<table>
<thead>
<tr>
<th>Property combinations</th>
<th>Attestations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. VSO / Pr / NG / NA</td>
<td>19</td>
</tr>
<tr>
<td>2. VSO / Pr / NG / AN</td>
<td>5</td>
</tr>
<tr>
<td>3. VSO / Pr / GN / AN</td>
<td>1</td>
</tr>
<tr>
<td>4. VSO / Pr / GN / NA</td>
<td>0</td>
</tr>
<tr>
<td>5. VSO / Po / NG / NA</td>
<td>0</td>
</tr>
<tr>
<td>6. VSO / Po / NG / AN</td>
<td>0</td>
</tr>
<tr>
<td>7. VSO / Po / GN / AN</td>
<td>0</td>
</tr>
<tr>
<td>8. VSO / Po / GN / NA</td>
<td>0</td>
</tr>
<tr>
<td>9. SVO / Pr / NG / NA</td>
<td>21</td>
</tr>
<tr>
<td>10. SVO / Pr / NG / AN</td>
<td>8</td>
</tr>
<tr>
<td>11. SVO / Pr / GN / AN</td>
<td>3</td>
</tr>
<tr>
<td>12. SVO / Pr / GN / NA</td>
<td>1</td>
</tr>
<tr>
<td>13. SVO / Po / NG / NA</td>
<td>0</td>
</tr>
<tr>
<td>14. SVO / Po / NG / AN</td>
<td>2</td>
</tr>
<tr>
<td>15. SVO / Po / GN / AN</td>
<td>6</td>
</tr>
<tr>
<td>16. SVO / Po / GN / NA</td>
<td>11</td>
</tr>
<tr>
<td>17. SOV / Pr / NG / NA</td>
<td>4</td>
</tr>
<tr>
<td>18. SOV / Pr / NG / AN</td>
<td>0</td>
</tr>
<tr>
<td>19. SOV / Pr / GN / AN</td>
<td>1</td>
</tr>
<tr>
<td>20. SOV / Pr / GN / NA</td>
<td>0</td>
</tr>
<tr>
<td>21. SOV / Po / NG / NA</td>
<td>7</td>
</tr>
<tr>
<td>22. SOV / Po / NG / AN</td>
<td>0</td>
</tr>
<tr>
<td>23. SOV / Po / GN / AN</td>
<td>28</td>
</tr>
<tr>
<td>24. SOV / Po / GN / NA</td>
<td>24</td>
</tr>
</tbody>
</table>

(Boolean, yes/no) attributes on the dimensions of adpositions, adjective–noun, and genitive–noun, he comes up with a typology classifying the languages of the world into $3 \cdot 2 \cdot 2 \cdot 2 = 24$ logically admissible types. Basing his analysis on 142 (groups of) languages of wide genetic and areal coverage, he notes that 9 out of the 24 logically admissible types are actually unattested. This is seen in Table 1, which summarizes Greenberg’s Appendix II. The disallowed Types are 4, 5, 6, 7, 8, 13, 18, 20, and 22.

In his paper, Greenberg proposes 28 non-statistical and statistical ordering universals, basing his analyses, in addition to Appendix II, on his 30-language sample (given in Appendix I) which contains information also on other ordering properties such as noun–relative clause, auxiliary verb–main verb, etc. Here, we will be interested only in the non-statistical universals pertaining to the proposed four-dimensional typology, i.e., the ones referring to the ordering
of verb–subject–object, adpositions, adjective–noun, or genitive–noun, since only these are relevant to our task of describing the typology.

Greenberg managed to find two exceptionless universals holding in Appendix II: his Universal 3 (“Languages with dominant VSO order are always prepositional”)\(^2\) and his Universal 5 (“If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun”).

The goal of positing universals, which Greenberg was in pursuit of, is to define all and only the attested types in a typology. Operating in conjunction with one another, a set of universals should predict which types occur and which do not. Thus, a word order type would be attested if none of its word orders violated any of the conjoined universals; otherwise, this type would be unattested. Or, putting it differently, any word order co-occurrence type would be attested if it satisfied the logical conjunction [Universal\(_1\) & Universal\(_2\) \& \ldots \& Universal\(_n\)], which means (recalling the definition of logical conjunction) that each of the conjuncts, Universal\(_1\), Universal\(_2\), etc., should be individually satisfied; if, in contrast, a type has some orderings that do not satisfy even one of the conjuncts, then this type ought to be unattested.

How successful has Greenberg been in describing his typology, and in doing it in the simplest possible way? Greenberg’s concern for simplicity is manifest in words like “In a certain sense we would prefer to have as few universals as possible, not as many. That is, we would like to be able to deduce them from as small a number of general principles as possible” (1966: 75). Though the principles he had in mind would perhaps be of a more abstract and explanatory nature than those needed for a typology’s description in terms of universals, we may safely assume that the simplest descriptive solutions would also be a worthy goal for him.

In order to evaluate Greenberg’s proposal, below we list, in symbolic form,\(^3\) his two universals, giving in parentheses the types excluded by each of them:

\[
\begin{align*}
\text{Universal 3:} & \quad \text{VSO} \rightarrow \text{Pr} & (\text{Types 5, 6, 7, 8}) \\
\text{Universal 5:} & \quad \text{SOV} & \& \text{NG} \rightarrow \text{NA} & (\text{Types 18, 22})
\end{align*}
\]

\(^2\) As a point of fact, this universal turned out to have a counterexample, viz., Papago. However, since this language was pointed out to Greenberg only after the completion of his article, its actual existence should be ignored as it will be irrelevant to our purely methodological discussion attempting to assess how successful Greenberg has been in describing the data at his disposal. Cf. also Note 1.

\(^3\) We use the following common logic notations: \(\neg\) (negation), \(\rightarrow\) (implication), \& (conjunction), \(\lor\) (disjunction), \(\equiv\) (equivalence). In what follows, we shall state universals only in symbolic form, rather than also verbalize them, as their interpretation will be obvious to a linguistic audience conversant with typological and universals research.
(It is easy to see in Table 1 that these universals block exactly these types, and no others.) If we compare all the types excluded by the universals with all the unattested ones, we see that Types 4, 13, and 20 are not ruled out by any of the generalizations even though they are unattested. Greenberg’s generalizations did not constitute a descriptively adequate account of the data in Appendix II and therefore the question of whether his account is minimal or not simply does not arise.

In his important book Hawkins (1983) extends and slightly corrects Greenberg’s typology in Appendix II, using the data from his Expanded Sample of 336 (groups of) languages. Table 2 summarizes Hawkins’ Expanded Sample (1983: 288).

Hawkins preserves the basic structure of Greenberg’s typology in keeping the same dimensions and the same attributes on these dimensions, with one exception. Thus, instead of the attribute VSO, Hawkins introduces the attribute
Vladimir Pericliev

V-1, basically in order to admit both VSO languages and VOS languages, the latter discovered by Keenan (1978) (for details, cf. Hawkins 1983: 55). (As seen in Table 2, no explicit provision is made for the object-first languages found by Derbyshire & Pullum (1981).) This typology, again, has 9 unattested word order types, which, however, are slightly different from those of Greenberg (cf. Table 1); non-occurring here are Types 4, 5, 6, 8, 13, 14, 18, 20, and 22.

Hawkins (1983) repeatedly states that the major goal of the linguist is to describe the data most simply in terms of implicational universals. Thus, for example, he writes that “the purpose of a set of implicational universals, operating collectively, is to define all and only the attested word order co-occurrences in the most revealing, and SIMPLEST, manner” (1983: 29; emphasis added). He declares that Greenberg has only been partly successful in this task (Hawkins 1983: 27), and addresses the problem head-on in Section 3.2 of his book.

Below is the set of implicational universals he found, with the excluded types given in parentheses:

Universal (I) SOV → (AN → GN) (Types 18, 22)
Universal (II) V-1 → (NA → NG) (Types 4, 8)
Universal (III) Pr & ~SVO → (NA → NG) (Types 4, 20)
Universal (IV) Po → (AN → GN) (Types 6, 14, 22)

Hawkins notes that two remaining unattested types, viz. Types 5 and 13, are not forbidden by his universals (I)–(IV). Apparently assuming that there ex-

---

4. Henceforth we preserve Arabic numerals for the enumeration of Greenberg’s universals related to Appendix II and Roman numerals for Hawkins’ universals related to the Expanded Sample. The universals we posit will have a prime ′ added to an Arabic or Roman numeral depending on whether they describe Appendix II or the Expanded Sample, respectively. As regards the symbolic formulation of Hawkins’ universals, we may note that his “complex” universals of type A → (B → C) are in fact logically equivalent (by an equivalence known in logic as exportation) to A & B → C; the universals A → (B → (C → D)) are equivalent to A & B & C → D, etc. The latter symbolization seems more intuitive to us for the following reasons. First, it can be more naturally verbalized than the former statement (cf. “If A and B, then C” vs. “If A and B, then C”). Secondly, the truth tables for the latter formulae containing & (and) are more immediately evident than in the former notation, viz., they are the same as for the two-termed implication A → B, in which the antecedent, A, is a conjunction of atomic terms. And, finally, since in the latter symbolization a conjunction of terms denotes the antecedent of an implication, and since a conjunction (like disjunction or equivalence) is commutative, it is obvious that the permutation of these terms is logically immaterial: A & B & C → D is equivalent to B & A & C → D, etc. In this context, it makes no sense from a logical (or linguistic) perspective to speak of an “ultimate antecedent”, as it happens a number of times in Hawkins (1983). The greater transparency of a complex universal being stated with one conjunctive antecedent is also noted by Dryer (1997: 141). Despite these considerations, however, we shall keep Hawkins’ original symbolization as it has been widely used in the linguistic literature throughout the years.
ists no set of exceptionless universals that can fully describe the typology, including Types 5 and 13, he leaves them aside, suggesting that statistical generalizations should be invoked for explanation of their non-occurrence. Thus, Hawkins writes that these types “will be predicted by the distributional principle in Chapter 4 to be rare or nonoccurring in a sample this size” (Hawkins 1983: 69). Understandably, as with Greenberg, no attempt is made to justify the proposed universals as the simplest account.

Summarizing the discussion so far, we may conclude that while Greenberg’s analysis of the data in his Appendix II has only been partially successful in that he failed to exclude three unattested types, Hawkins’ account of his Expanded Sample is not fully satisfactory either because it still fails to exclude two unattested types. In both cases, the proclaimed goal of descriptive simplicity has been compromised owing to the failure of both authors either to find several complete solutions – a necessary prerequisite for selecting the simplest one – or alternatively, to demonstrate that some solution is unique and hence the simplest.

These objections are valid only if there do exist sets (conjunctions) of non-statistical universals that define all and only the attested types in both Table 1 and Table 2 in a simplest way. In the next section, we postulate a descriptive principle stating that for typologies such as those in Table 1 and Table 2 such minimal accounts indeed do exist. This principle thus allows us to claim in an a priori way that both Greenberg’s and Hawkins’ analyses are not fully adequate, even if we do not know what the correct analyses actually are.

3. A descriptive principle in typology

In a most revealing article, Greenberg (1978) brings to the attention of linguists some logical properties of the different types of typologies used in linguistics. He discusses typologies in terms of the number of their dimensions (one- or multi-dimensional typologies), the attributes on these dimensions (categorical or numerical), and, for the case of one- and two-dimensional typologies, describes the logical form to which they correspond (e.g., one-dimensional typologies correspond to unrestricted universals, two-dimensional typologies to unilateral or bilateral implications, etc.). Greenberg does not enter into analysis of the logical expressions generated by typologies of more than two dimensions.

Before considering the question of describing arbitrary typologies in its full generality, let us limit our attention to typologies having only binary (i.e., Boolean, yes/no) attributes on all their dimensions.

From Greenberg’s discussions, as well as from elementary knowledge of mathematical logic, it is clear that an \( n \)-dimensional typology \((n \geq 1)\) with binary attributes for these dimensions is in fact an \( n \)-place \((n\text{-argument})\) truth
function (truth table) in propositional logic. Thus, the $n$ dimensions of the typology correspond to the $n$ places (= arguments) of the truth function, the binary yes/no attributes of the typology correspond to the values T(ue)/F(alse) the truth function’s arguments take, and the attested vs. unattested types in a typology correspond to the two possible values T(ue)/F(alse) of the truth function. An $n$-dimensional typology with binary attributes defines all logically possible types whose number is $2^n$, and $2^n$ is exactly the number of the rows in a truth table, which result from all logically possible distributions of the values T/F of all the arguments of the truth function.

Now, let us state some relevant facts from propositional logic. For our discussion, we need only mention them rather than go into the details of how they may be formally proven in mathematical logic (the interested reader is referred, e.g., to Mendelson 1963: Chapter 1, or Quine 1965: Chapter 1).

(i) Every truth function can be generated by some propositional formula.

(ii) Every propositional formula can be represented as a CONJUNCTIVE NORMAL FORM (CNF). A CNF is a propositional formula of the type $C_1 \& C_2 \& C_3 \ldots \& C_n$, in which every conjunct $C_i$ is a disjunction of atomic propositions or their negations. Examples of CNFs are: $(A \lor B \lor \sim C) \& (A \lor D)$; $(A \lor B) \& (C \lor D)$; $A$, etc.

(iii) Every compound (= non-atomic) propositional formula can be expressed by means of atomic propositions bound only by one minimal pair of logical connectives: the pair negation and implication ($\sim$ and $\rightarrow$), the pair negation and conjunction ($\sim$ and $\&$), or the pair negation and disjunction ($\sim$ and $\lor$); alternatively, several of these connectives may be used.

Now, given that an $n$-dimensional typology with binary attributes is equivalent to an $n$-place truth function, it will follow by (i) that it can be generated by some propositional formula (= a (compound) universal). This (compound) universal, in turn, can be expressed as a Conjunctive Normal Form (by (ii)). And, finally, each individual conjunct in this (compound) universal in CNF, if compound, may be represented, among other alternatives, by an implicational expression possibly containing negations, or optionally, other connectives as well (by (iii)).

In purely linguistic terms, the argument above amounts to the following: Any $n$-dimensional typology with binary attributes can be described by a non-statistical (compound) universal, or what is the same, by a conjunction of universals, of the form:

$$[\text{Universal}_1 \& \text{Universal}_2 \& \text{Universal}_3 \ldots \& \text{Universal}_n]$$

5. A truth function of $n$ arguments is any function of $n$ arguments which takes the truth values True or False, its arguments also taking the same truth values.
Besides, any of these conjoined universals, if compound, will consist of atomic propositions linked by disjunctions and negations, and can be represented either in implicational form (replacing the disjunctions and negations with implications and negations, and possibly some other connectives) or in some other logical form – an option that is less popular in linguistics.6

These considerations allow us to postulate the following “existence” principle:

For any linguistic \(n\)-dimensional typology with binary attributes, there exists at least one set (conjunction) of non-statistical universals of implicational (or some other) form which generates this typology (i.e., describes all and only its attested types).

Let us now turn to the typologies given in Table 1 and Table 2. The only difference between these typologies and one having binary attributes is the non-binary nature of their dimension for Verb Order. This dimension has three attributes, viz., SVO, SOV, and VSO (or V-1), instead of only two; all other dimensions are binary. But obviously we can split up this 3-attribute dimension into three 2-attribute dimensions, SVO, SOV, and VSO (or V-1), which are now binary, as each language type will either have or lack any one of the mentioned orders. More generally, any \(n\)-dimensional typology with non-binary attributes can be transformed into one with binary attributes by increasing the number of dimensions.7

Two specific types of \(n\)-dimensional typologies deserve special mention at this point. The first is the one in which ALL the types are actually attested, and the second is the one in which NONE of the types are attested. Typologies with no attested types can be regarded as non-occurring in linguistic practice, and hence ignored. The reason why such typologies are useless is simply that the

---

6. The details of a conversion of one logical formula into another equivalent formula will be of no great interest to a typologist. It suffices for the purposes of the present article that only the basic idea of our argument be grasped, or indeed only its ultimate result. Therefore, a single simple example of such a conversion will have to satisfy the more logically-minded typologist. Thus, let Universal\(_1\) from the conjunction of universals \([\text{Universal}_1 \& \text{Universal}_2 \& \text{Universal}_3 \& \ldots \& \text{Universal}_n]\) be the disjunction \(A \lor B\). We want to get an implicational formula from this disjunction. First, we can substitute \(A\) by its equivalent formula \(\neg(\neg A)\) (by the law of double negation, \(A \equiv \neg(\neg A)\)), and thus obtain \(\neg(\neg A) \lor B\). This result then is convertible into the implication \(\neg A \rightarrow B\) by the “material conditional” \(A \rightarrow B \equiv \neg A \lor B\).

7. It is important to realize that this reduction move is only a mental operation in the construction of our argument and therefore does not involve, in any literal sense, the transformation of non-binary to binary typologies. As a consequence, any questions naturally arising in linguistic feature theory, such as intuitiveness vs. unintuitiveness of binary typologies, the advantages and disadvantages of particular ways the reduction is achieved, etc. are completely beside the point in the present context. All that counts is the possibility of the reduction. Our illustrative examples in the next section are in fact both non-binary typologies.
properties they employ to classify languages must be irrelevant to natural languages, if indeed no language is either positively or negatively specified with respect to these properties. (An example would be a typology attempting to classify languages, say, on the basis of the properties “has feathers” and “can fly”.) While typologies with no attested types have not been proposed in the literature, ones with all types attested have been. They state that any combination of linguistic properties is realized, which means that no exceptionless generalization can formulated describing such typologies. The reason why this is so can easily be understood by appealing to the familiar fact that any universal FORBIDS some co-occurrence of linguistic properties, or some type(s), but, by definition, no type is forbidden in the typologies at issue. In facing such typologies with no forbidden types, which do not allow the formulation of non-statistical universals, the linguist has to look for statistical universals, if some significant statistical correlations between the languages’ attributes in the typology are discernible. (A statement to the same effect, but for the more specific case of two-dimensional typologies, is made by Greenberg when he says that for tetrachoric tables with four pluses “no exceptionless generalization is possible” (1978: 54).)

The above considerations, pertaining to the reducibility of typologies to binary ones and to generalizations corresponding to typologies with all/none forbidden types, allow us to reformulate our previous version of the existence principle as follows, where the phrase “n-dimensional typology with binary attributes” is replaced by “n-dimensional typology such that it has some attested and some unattested type(s)”:

For any linguistic n-dimensional typology such that it has some attested and some unattested type(s), there exists at least one set (conjunction) of non-statistical universals of implicational (or some other) form which generates this typology (i.e., describes all and only its attested types).

This is a fundamental descriptive principle in typology since it tells us that any typology that has some attested and some unattested types does have some description in terms of a set of exceptionless universals covering all and only the attested language types, even if we do not know exactly what this set of universals might be. The heuristic value of this “existence” principle is self-evident. Our knowledge that a solution of some pre-specified format exists is, in the first place, a good incentive for initiating the search for this solution. And, secondly, it will direct this search by precluding the possibility of someone trying to pass for the correct solution an object that deviates from the pre-specified solution’s format.

It may be worth noting at this point, following Greenberg (1978: 41–49), that not all typological work necessarily employs typologies (≡ classifications) of the usual sort, where there is a finite number of mutually exclusive types (≡
classes), and every language relevant to the typology falls under one and only one type (class). As a result, not all typological descriptions are summarizable in terms of universals. Thus, typological characterization on the basis of continuous numerical attributes – for example, the use of morpheme–word ratio as a measure of the degree of typological synthesis (Greenberg 1954) – does not divide languages into mutually exclusive classes but orders them on a continuum. Indeed, if each language is assigned a number corresponding to its morpheme–word ratio over a sample of texts from this language, the languages do not form classes but are ordered on a numerical continuum. This allows us only to register, say, that Inuit is more synthetic than Vietnamese, and German lies somewhere between them, but does not let us set up classes to which these languages belong. As Greenberg (1978) notes, in this case the associated generalizations involved would be statistical measures of central tendency (e.g., averages and medians, dispersion, etc.), rather than (implicational) universals. Nonetheless, if desired, continuous numerical attributes can be reduced to categorical (nominal) attributes by defining number intervals. This reduction results in normal typologies that obey the existence principle stated above. For example, introducing the three morpheme–word ratio intervals \([< 2.00]\), \([2.00 – 2.99]\), and \([\geq 3.00]\), we introduce a dimension with three categorical attributes (corresponding to analytic, synthetic, or polysynthetic language, respectively). This dimension, possibly along with some further linguistic properties, can then be used to define an \(n\)-dimensional typology for which we can assess whether or not it is describable in terms of non-statistical universals in accordance with our existence principle.

A remark is in order on the simplicity of solutions to typologies. Our descriptive principle states that there would be at least one solution, i.e., one general description, to a typology, comprising a set of (implicational) universals. If we can demonstrate that a solution is unique, then this solution is clearly the simplest. In cases where there are alternative solutions, it is natural to regard the set(s) of universals with the smallest cardinality (= size) as simplest (e.g., a set of three universals is simpler than a set of four universals). Thus, in effect, for any typology of the type we discuss, there exists one or more simplest solutions.

Insofar as Table 1 and Table 2 do not belong to the class of typologies in which all or none of the logically possible types are attested, they will have minimal accounts in terms of non-statistical universals.

We are not aware of our descriptive principle having been previously stated in the linguistic literature. In any case, it seems unlikely that it was familiar to Greenberg or Hawkins (at the time of writing), for if it had been, they would undoubtedly have found a comprehensive account of their typologies. In particular, had this principle been known to Hawkins, it would have saved him the need to relegate Types 5 and 13 to his (statistical) distributional principle.
even though there were universals of the non-statistical sort favored by him that would have been sufficient to do the job.

4. A method for finding the minimal description(s) of a typology

It is good to know that a (simplest) solution to a problem exists but it is even better if you know how to find it. In this section, we propose a method of finding the minimal description(s) of a typology in terms of a set of implicational universals. We illustrate our approach on Hawkins’ typology (cf. Table 2).

Our method comprises the following steps:

Step 1. Find all logically nonequivalent implications holding over attested types; then associate each of these implications with the type(s) it forbids.

In order to find a minimal set of universals describing a typology we first need to state all universals that are valid for the data. To avoid the proliferation of generalizations potentially discoverable in the data, we may limit ourselves to finding only logically nonequivalent implications, as any generalization from a set of logically equivalent generalizations makes the same claim as any other.

Two propositions (= universals) \( P \) and \( Q \) are said to be logically equivalent if when \( P \) is true \( Q \) is also true, and when \( Q \) is true \( P \) is also true; otherwise they are logically nonequivalent. Such equivalence between two propositions (universals) can be ascertained by drawing their truth tables and checking whether they have identical truth values in each row of their truth tables, or alternatively, by showing that \( P \) is convertible into \( Q \) by some known tautology (= law of logic).

As a familiar example, consider the universal \( \text{SOV} \rightarrow (\text{NG} \rightarrow \text{NA}) \), which is equivalent to Hawkins’ universal (I), viz. \( \text{SOV} \rightarrow (\text{AN} \rightarrow \text{GN}) \). Their equivalence can be shown, using the logic law of contraposition: \( P \rightarrow Q \equiv \sim Q \rightarrow \sim P \). Thus, under the assumption made in the typologies investigated of a basic word order for the adjective and the noun, and for the genitive and the noun, it is obvious that \( \text{AN} \) is the negation of \( \text{NA} \) (i.e., \( \text{AN} \equiv \sim \text{NA} \)), and \( \text{GN} \) is the negation of \( \text{NG} \) (i.e., \( \text{GN} \equiv \sim \text{NG} \)). From these facts it is easy to see how the former universal is derivable from that of Hawkins’ by the appropriate substitutions. In an analogous manner, one can derive Hawkins’ universal (I) from the former (for a linguistic discussion of contraposition, cf., e.g., Croft 1990: 49). For our descriptive purposes, we need to keep only one of these redundant universals.

Below we list nine logically nonequivalent implications that we found to hold for Table 2.\(^8\) In their formulation we use (explicitly) only the connec-

\(^8\) To avoid the proliferation of redundant implications found, also excluded from this list are im-
Economy in formulating typological generalizations

Each of the implications is associated with the excluded co-occurrence types:

Universal (I')  SOV → (AN → GN)  (Types 18, 22)
Universal (II')  V-1 → (NA → NG)  (Types 4, 8)
Universal (III')  Pr → (NA → (GN → SVO))\(^9\)  (Types 4, 20)
Universal (IV')  Po → (AN → GN)  (Types 6, 14, 22)
Universal (V')  NA → (V-1 → Pr)  (Types 5, 8)
Universal (VI')  NG → (V-1 → Pr)  (Types 5, 6)
Universal (VII')  NG → (SVO → Pr)  (Types 13, 14)
Universal (VIII')  NG → (Po → SOV)  (Types 5, 6, 13, 14)
Universal (IX')  GN → (NA → (SOV → Po))  (Type 20)

Our universals (I')–(IV') coincide with Hawkins' set (I)–(IV).

**Step 2.** Associate with any unattested type the universal(s) which exclude that type.

This step is unproblematic once we have available all the implications found at Step 1; see Table 3 for Hawkins’ Expanded Sample.

**Step 3.** Form a set \(S\) whose members are all sets consisting of the alternative universals excluding a type (i.e., the sets in the right-hand column of Table 3), and then find the minimal set cover of \(S\).

The set \(S\) will have as its members all nine sets of universals in Table 3, i.e., \(S = \{(II', \ III'), (V', VI', \ VIII'), (IV', \ VI', \ VIII'), (II', \ V'), (VII', \ VIII'), (V', \ VII', \ VIII'), (I'), (III', IX'), (I', \ IV')\}\).

Plications that logically follow from stronger implications, such as the four-termed universal \(Po → (SOV → (AN → GN))\) which logically follows from Hawkins’ three-termed universal (I), \(SOV → (AN → GN)\). In general, a proposition (universal) \(P\) is said to logically imply another proposition \(Q\) if, when \(P\) is true, \(Q\) is also necessarily true. To see why Hawkins’ universal logically implies the other universal, we first note that Hawkins’ universal is equivalent (identical) to the consequent of the latter implication. Denoting Hawkins’ universal \(SOV → (AN → GN)\) by \(H\), we therefore need to show that whenever \(H\) is true, \(Po → H\) is also true. That this is indeed the case follows from the fact that when the consequent of an implication is true, the whole implication is also necessarily true. The superfluosness of universals that are logically implied by stronger universals has long been noted in the typological literature. The standard example is the (unrestricted) universal “All languages have oral vowels” which logically implies the (implicational) universal “If a language has nasal vowels, then it also has oral vowels”. Cf., e.g., Howard (1971), Greenberg (1978: 50–51), Comrie (1981: 18).

9. This formulation is equivalent (by contraposition and exportation) to Hawkins’ Universal (III) \(Pr & ~SVO → (NA → NG)\), which can also be seen from the same set of types they both rule out, viz. Types 4 and 20.
Table 3. Unattested types and universal(s) which exclude them

<table>
<thead>
<tr>
<th>Excluded type</th>
<th>Universals excluding the type</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>(II', III')</td>
</tr>
<tr>
<td>5</td>
<td>(V', VI', VIII')</td>
</tr>
<tr>
<td>6</td>
<td>(IV', VI', VIII')</td>
</tr>
<tr>
<td>8</td>
<td>(II', V')</td>
</tr>
<tr>
<td>13</td>
<td>(VII', VIII')</td>
</tr>
<tr>
<td>14</td>
<td>(V', VII', VIII')</td>
</tr>
<tr>
<td>18</td>
<td>(I')</td>
</tr>
<tr>
<td>20</td>
<td>(III', IX')</td>
</tr>
<tr>
<td>22</td>
<td>(I', IV')</td>
</tr>
</tbody>
</table>

A cover of a set S is called another set C which contains at least one member from each of the sets that are members of S. That is, the cover C in the above case should contain at least one member from the first set (II', III'), at least one member from the second set (V', VI', VIII'), at least one member from the third set (IV', VI', VIII'), and so on for all the nine sets of universals which are members of S. We note that a cover must not contain redundant elements. It will be clear from this definition that finding the cover of the set S will contain the universals needed to exclude all the nine unattested types. A minimal cover of a set S is the cover C, having the smallest cardinality (= the smallest number of members, composing C). The minimal cover of S will thus yield the smallest number of universals that can block all the non-attested types.

Computing the minimal covers for the set S, we get the following three minimal sets of universals, each consisting of four universals which describe all and only the attested types in Table 2:

Account 1: Universals [I' & II' & III' & VIII']
Account 2: Universals [I' & II' & VIII' & IX']
Account 3: Universals [I' & III' & V' & VIII']

There also exist five other, non-minimal, accounts, consisting of 5 and 6 universals, as follows:

Account 4: Universals [I' & II' & VI' & VII' & IX']
Account 5: Universals [I' & III' & IV' & V' & VII']
Account 6: Universals [I' & II' & III' & VI' & VII']
Account 7: Universals [I' & III' & V' & VI' & VII']
Account 8: Universals [I' & II' & IV' & V' & VII' & IX']

For example, account 1 uses the first three of Hawkins’ universals plus Universal (VIII') ruling out, among others, the non-attested Types 5 and 13 that
Hawkins had problems with. The correctness of the rest of the solutions is readily testable against the data from Table 2 and we leave this exercise to the reader. We may note that Hawkins’ Universal (IV) does not actually figure in any one of the three simplest accounts. The reason is simply that it does not add any further information to what is already contained in any of these three accounts.

We should emphasize that, since a cover of a set does not contain redundant members (i.e., universals that are logically equivalent to or logically implied by other universals), both the minimal and non-minimal accounts above contain all and only the universals needed to describe the typology in question. That is, adding a universal to any of the accounts would result in redundancy whereas removing a universal would lead to a failure to describe the typology. This is the reason why a solution that includes, say, all nine universals found to hold in the Expanded Sample would not be a correct one; more precisely, its shortcoming will be the presence of superfluous universals.

We may now look at the accounts of Greenberg’s typology in his Appendix II. At Step 1 of our method, we found the following implicational universals listed below alongside with the co-occurrence types each excludes:

| Universal (1’) | VSO → Pr (Types 5, 6, 7, 8) |
| Universal (2’) | SOV → (NG → NA)\(^{10}\) (Types 18, 22) |
| Universal (3’) | VSO → (NA → NG) (Types 4, 8) |
| Universal (4’) | AN → (NG → (Po → SVO)) (Types 6, 22) |
| Universal (5’) | GN → (NA → (Pr → SVO)) (Types 4, 20) |
| Universal (6’) | GN → (NA → (SOV → Po)) (Type 20) |
| Universal (7’) | NA → (NG → (Po → SOV)) (Types 5, 13) |
| Universal (8’) | NA → (NG → (SVO → Pr)) (Type 13) |

The following two are the minimal sets of universals that define Appendix II. They comprise four universals:

- **Account 1**: [1’ & 2’ & 5’ & 7’]
- **Account 2**: [1’ & 2’ & 5’ & 8’]

There are also two non-simplest accounts each comprising five universals:

- **Account 3**: [1’ & 2’ & 3’ & 6’ & 7’]
- **Account 4**: [1’ & 2’ & 3’ & 6’ & 8’]

Universals (1’) and (2’) correspond to Greenberg’s Universals 3 and 5, respectively. Hence both generalizations proposed by Greenberg should figure

---

10. This universal is equivalent (by exportation) to Greenberg’s SOV & NG → NA as will be clear from Note 4.
in both the simplest and non-simplest accounts of his typology, as seen from accounts 1–4.

Concluding this section, a methodological remark is in order. As seen from our analyses above, both Appendix II and the Expanded Sample allow more than one simplest description (under our definition of “simplest” as referring to the minimal number of universals in an account of a typology). None of the simplest solutions can be considered “better” than any other from a purely descriptive point of view as all of them are equally empirically adequate. If simplicity is the seal of truth, as the Latin saying quoted at the beginning of this article has it, then we have to concede that there exists in this case more than one solution marked with this seal.

5. The MINTYP program
The reader will have noticed by now that the execution of our method is not straightforward as far as Step 1 and Step 3 are concerned. The tasks defined by these steps are indeed quite complex computationally and hence very difficult to perform manually even for the relatively small typologies we are considering.

In order to execute the method proposed we have implemented the computer program MINTYP (running in SICStus Prolog under Windows). MINTYP accepts as input information such as that in Tables 1 and 2, i.e., the definitions (= co-occurrences) of all logically admissible types, alongside information about the number of languages conforming to each type. That is, the typologies are inputted to the system in the form in which they are commonly stated in linguistics.

In essence, MINTYP links two basic modules, one for discovery of universals and the other for finding minimal covers. Both have been previously built for the programs UNIV(ersals) and KINSHIP, respectively.

KINSHIP is capable of finding all minimal (or non-minimal) componential models of a kinship system, given as input the kin terms of a language with their attendant kin types. For a linguistic discussion, cf. Pericliev & Valdés-Pérez (1998a, b); the computational machinery, which is very general and hence usable for other linguistic (and non-linguistic) tasks, is described in detail in Valdés-Pérez & Pericliev (1997) and Valdés-Pérez et al. (2000).

UNIV (Pericliev 1999a) is a system that can discover logical patterns from data including both statistical and non-statistical universals as well as estimate their statistical significance (for one approach to the latter issue, see Valdés-Pérez & Pericliev 1999).11

---

11. It would be interesting to note that recently UNIV was extended with a text generation module, AUTO (AUthoring TOol), so that given a description of a set of languages in terms of
MINTYP uses the two modules to execute the steps of the method outlined above. At step 1, the UNIV module is run on data compiled only from the attested types of the inputted typology to find (non-redundant) implications of increasing length. Next, a subroutine is run on the data from the entire typology (i.e., attested plus non-attested types) to discover the types that are counterexamples to the implications found. The latter process identifies the forbidden types, which are associated with each implication forbidding them. Step 2, comprising the association of unattested types with the universal(s) which exclude them, is unproblematic. And finally, at Step 3, the minimal cover module from the KINSHIP program is applied to the set consisting of all sets of universals ruling out a type to yield the minimal account(s) of a typology. (This module can optionally produce all alternative sets of universals that define the typology, not only the simplest one(s).)

The algorithms we use for finding all (non-redundant) universals and minimal set covers are complex. They are of more interest to computer science than to linguistic typology, and thus they are beyond the scope of this article. The essential point for typologists is that the task of finding either a guaranteed-simplest or a non-simplest account of a typology – whether using this particular, or some other, method – is complex and would generally require a computational tool to perform it. As we have pointed out earlier, for some data-intensive tasks such as the one discussed here the difference between using and not using a computer is not just a matter of saving time and effort, but often a matter of finding and not finding a consistent solution at all. (For a general discussion of the advantages of computer-aided knowledge-acquisition in linguistics, cf. Pericliev 1999a.)

6. Conclusion

The contributions of this paper may be summarized as follows. We proposed a principle asserting that, for any typology with some unattested types, there exists a (simplest) set of non-statistical universals of some logical form (e.g., implications) that define all and only its attested types. This was made possible...

some properties (currently only word order properties), as well as the discoveries of a human agent on the same data, the UNIVAUTO system can compare the human and the machine discoveries, and if judged worthy by the system, write an article in English on its discoveries. Running UNIVAUTO on Greenberg’s 30-language sample, with queries requiring non-statistical or statistical universals of type A → B to be found, the system both exhibited some infelicities in Greenberg’s universals (1966) and found many novel generalizations of these types. These two sessions with AUTO resulted in texts that were submitted for publication with no further human editing (Pericliev 1999b, 2000; the first paper has a postscript giving a brief description of UNIVAUTO). Thus, UNIVAUTO has produced the first scientific articles ever to be generated by a computer.
by building on the insightful logical analysis of linguistic typologies by Greenberg (1978) and linking his analysis to certain relevant results from propositional logic. The postulated principle allowed us to judge both Greenberg’s and Hawkins’ descriptions as not fully descriptively adequate, even without knowing what exactly the descriptions of their typologies would be. We also suggested a method of discovering the (minimal) description(s) of a typology in terms of implicational universals and briefly outlined the computer program MINTYP that executes this computationally costly task. Running the program on the data from Greenberg’s Appendix II and Hawkins’ Expanded Sample, we found the minimal (as well as the non-minimal) sets of universals defining all and only the attested types in these typologies and showed that each set must consist of at least four implicational universals. We also noted the existence of alternative simplest sets of universals defining the typologies in question.

The set of universals defining a typology may be viewed either as a set of grammar rules generating attested co-occurrences or as mere facts of co-occurrence that need to be explained by some theory. Under both interpretations, having a most economical description at our disposal might be a virtue. In the first case, it is generally preferable to have as few grammatical rules as possible; in the second case, again, the availability of a smaller number of facts to explain is generally preferable to having a much greater number of facts to explain. Although we have focussed on finding guaranteed-simplest solutions, it will be clear from our discussion that finding ANY consistent description of a typology, be it the simplest or not, would involve a similar method to the one we described here.

Finally, a word of caution against a possible (mis)interpretation of our results as implying the superfluousness of statistical universals for the description of typologies. Indeed, one may feel tempted to claim that since for many typologies there exist set(s) of non-statistical universals defining these typologies, there is no place for statistical universals in this enterprise. The debate in favor or against statistical universals is a complex matter in which the arguments of advocates of either position should be carefully weighed (cf., e.g., Hawkins’ 1983 defense of non-statistical universals, and Dryer’s 1997 of statistical ones). This is a task beyond the scope of this study. Here, we shall have to limit ourselves to a few remarks as to why our results should NOT be conceived as downplaying the role of statistical universals.

In the first place, and this is an obvious point, non-statistical universals cannot register prevailing co-occurrence tendencies or significant correlations that have a limited number of exceptions; such important facts about language are only accountable for in terms of statistical universals. Secondly, on inspecting increasingly larger databases it may eventually turn out that all of the logically possible types are actually attested; in this case, again, as mentioned earlier (Section 3), we ought to take recourse to statistical, rather than non-statistical,
Economy in formulating typological generalizations

universals for a description of this typology. Thus, for instance, it was found that Type 4 (V-1 & Pr & NA & GN) and Type 8 (V-1 & Po & NA & GN), previously believed to be non-existent, actually occur in languages such as Kilivila and Garawa, and Yagua and Guajajara, respectively (for a discussion, cf., e.g., Dryer 1991). If it is convincingly demonstrated that all types actually occur, we will be forced by these empirical data to replace the non-statistical descriptions of the four-dimensional typologies studied here with statistical ones. And, finally, since the ultimate aim of describing typologies in terms of universals is to provide data that can be subsequently deduced from, or explained by, higher-order principles or theories, there is definitely a place for statistical universals if they happen to fit better into these higher-order explanatory frameworks than non-statistical universals do.

Received: 1 December 2000 Revised: 13 April 2001

Correspondence address: Mathematical Linguistics Department, Institute of Mathematics and Informatics, bl. 8, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria; e-mail: peri@math.bas.bg

Acknowledgments: This work was partly supported by contract #I-813 (“Aspects of Machine Discovery and Applications to Linguistics”) with the Bulgarian Ministry of Education and Science. Some of the computational machinery used was developed under a grant #IIS-9819340 from the Division of International Programs of the US National Science Foundation (NSF).


References


12. An obvious implication of this point is that the only conclusive demonstration that some typology must be described in terms of statistical universals is showing that all its types are attested, NOT the piecemeal engineering of refuting any universal that happens to be posited on these data. The reason is simply that, following the latter strategy, one can never be sure that there are no non-statistical universals defining the typology after all.


