IX. Gebiete und Phänomene: Typologie / Fields and phenomena: typology

40. Quantitative methods in typology

1. Introduction

The principle of linguistic typology is that insight into the structure of human language can be obtained by classifying languages into types. The diversity and distribution of types helps us understand the possibilities and preferences of human language. The traditional conception of such types was holistic, meaning that the typology attempts to characterise a complete language as belonging to a particular type. At least since Greenberg (1963) a different, more reductionistic approach has arisen, in which only restricted domains of linguistic structure are classified into types, e. g. the kind of word order or the size of the phoneme inventory, to name just a few random examples (see Plank 2001 for a survey of 17th and 18th Century precursors of the reductionistic approach). It then becomes an empirical question whether the resulting typologies of different domains correlate with each other or not. By investigating such correlations in a sample of languages that are genealogically and areally independent from each other, an attempt is made to uncover regularities, or even universals, of linguistic structure.

In this approach, the usage of data from as many as possible languages is encouraged, but quantitative methods are not widely used. Of the 37 articles that appeared in the first six volumes of the journal *Linguistic Typology* (1997−2002) there are twenty that compare data from a wide variety of languages. However, only five of these twenty articles were based on some kind of representative sample of the world's languages and only five out of twenty (not necessarily the same) used some kind of quantitative analysis of the data. In fact only one article (viz. Siewierska 1998) actually presented frequencies as found in a stratified sample of the world's languages and only one other article (viz. Fenk-Oczlon/Fenk 1999) presented quantitative analyses on a cross-linguistic convenience sample of 34 languages. As far as the publications in *Linguistic Typology* are representative of the field of linguistic typology, this indicates that the usage of strict sampling procedures and quantitative analyses is not widespread.

The scarcity of usage of quantitative methods is also reflected in their somewhat unsophisticated application. Even so, I will survey the various quantitative approaches that have been used in the literature. In contrast to other surveys of quantitative methods in typology, like Altmann/Lehfeldt (1973) or Perkins (2001), I will not focus on measures and statistical tests that could be used in typology, but only discuss those methods that have actually been used and point out possible pitfalls with them. The present exposition will be organised along three main themes. First, in section 2, I will discuss various approaches to the problem of sampling. The central question here is which of the thousands of languages should be investigated in a typological study and what conclusions can be drawn from any such sample. Next, in section 3, the problem of establishing types will be discussed. In most contemporary typological investigations the types are defined qualitatively. In this section, I will summarise some quantitative methods to classify a language as belonging to a particular type. Finally, in section 4, the interpretation of typological data is examined from a quantitative point of view.

2. Sampling

2.1. Using data from many languages

When one intends to use data from a wide array of languages, the first question that arises is which languages one should investigate from among the 5.000 to 10.000 languages presently spoken. In most typological

studies, the set of languages chosen is a convenience sample, meaning that there is no *a priori* restriction on which languages might or should be included. Indeed, this is the best way to go for any exploratory study – and most typological investigations are still exploring the linguistic potential of human language. However, as soon as general patterns are observed, it is important to check such patterns in a more thoughtfully selected sample of the world's languages. This is the only way to assess the merit of any hypothesised generalisation about human linguistic structure. The question how to establish such a sample is the most widely discussed aspect of typological methodology, though this still means only about five original contributions (viz. Bell 1978; Dryer 1989; Perkins 1989; Rijkhoff *et al.* 1993; and Maslova 2000a).

The most widespread approach to sampling, as used throughout the social sciences, is to represent the diversity of phenomena using a stratified probability sample. A probability sample is, roughly spoken, a randomly chosen subset of the world's languages. By adding a stratification it is possible to delimit the influence of known biases. I will discuss various guidelines as proposed in the literature on how to compose a stratified probability sample in section 2.2. A general problem of probability samples is that the best they can do is to represent the actual world's languages, which are not necessarily the same as the possible human languages. Various reactions to this discrepancy will be considered in section 2.3.

The major form of criticism from the wider linguistic community to any kind of sampling is to point out errors of observation (i. e. "incorrect attribution of characteristics to languages", Bell 1978, 126). Typically, such criticism takes the form as found in Campbell *et al.* (1988) who react to Hawkins (1983) by "correcting some wrongly reported word-order patterns in certain languages […] We make no attempt to be exhaustive, but rather concentrate on languages of our experience and patterns that otherwise seem suspicious, i. e., are of a low frequency of occurrence in the sample" (Campbell *et al.* 1988, 210). It is of course of great importance to correct errors, but errors as such are not a major problem for a sample study. Errors will always be present and often they will neutralize each other (though the larger the sample, the greater the chance that some erratic residue will remain). A much more valid kind of criticism is to show that there is a consistent direction in the errors, leading to a systematic bias in the sample. The corrections as made by Campbell *et al.* (1988) do exactly the opposite: by focussing on a particular subset of the sample (viz. those languages they know well, and those languages that have an uncommon type in the original study by Hawkins) they induce a bias, thereby reducing the validity of the sample.

To be able to discuss the various approaches to sampling, a short terminological clarification is necessary. I will use the term *genus* for "genetic groups roughly comparable to the subfamilies of Indo-European, like Germanic and Romance" (following Dryer 1989, 267; cf. Bell 1978, 147; the term *family* is used for the same concept by Nichols 1992, 24). There appears to be rather general consensus about this notion, although nobody has been able to give a stricter definition than the one cited above. In an attempt to do so, a genus is sometimes equated with a group of related languages, which have maximally diverged for 3500 years (Bell 1978, 147; Dryer 1992, 84 n. 2). However, units that are called genera are often much less old – and sometimes much older (Nichols 1997, 362–363). Also note that looking at groups with a maximal divergence of 3500 years is not the same as looking at the number of languages 3500 years ago (Dryer 2000, 345–346). Besides *genus*, the term *stock* will be used for the maximal reconstructable unit, i. e. the highest node in a genealogical tree (cf. Bell 1978, 148; Nichols 1992, 25; the term *phylum* is used for the same concept by Perkins 1992, 128). The content of this notion of course highly depends on whose reconstruction one is inclined to believe.

2.2. Probability samples

The first extensive discussion of sampling techniques applied to linguistic typology is presented by Bell (1978). He strongly encourages the usage of a stratified probability sample and discusses many options of sample stratification. However, he only works out a stratification along genealogical lines in any detail. For each stock he estimates the number of genera. The number of languages sampled for each stock should be proportional to the number of genera in the stock (Bell 1978, 147–149). For example, Bell lists

12 genera for Indo-European, out of a total of 478 genera for the whole world, thus a sample should contain 12 / 478 = 2.5 % Indo-European languages.

The concept of a purely genealogical stratification has been perfected by Rijkhoff *et al.* (1993; 1998). Their method is designed to increase the probability of a rare type being represented in the sample. For each stock, they consider the complete structure of the genealogical tree to compute the *diversity value* (*DV*), relative to which the sample should be proportional. The formula to compute the *DV* for a particular stock is shown in (1). In this formula, L is the number of levels of the genealogical tree and N_x is the number of nodes at level x. The formula basically adds together the change in the number of nodes at each level ($N_x - N_{x-1}$), but it values the additions in the higher levels of each stock as more important than the additions in the lower levels (as expressed by the factor $L - x + 1/L$ in the formula). The other factor ($L_{\max} - x + 1/L_{\max}$) adds a kind of normalisation between stocks, limiting the *DV* of 'deep' stocks with many intermediate levels ($L_{\max}$ being the maximum number of intermediate levels of any tree in the world, in their case 16 from Niger-Kordofanian).

$$DV = \sum_{x=1}^{L} (N_x - N_{x-1}) \frac{(L - x + 1)}{L} \frac{(L_{\max} - x + 1)}{L_{\max}} \quad (1)$$

A particularly nice aspect of this approach is that it can be applied recursively within genealogical units to decide from which part of the tree a language has to be chosen. Rijkhoff *et al.* also propose to include at least one language for each highest node, an approach that leads to a diversity sample (Rijkhoff *et al.* 1993, 184–190; Rijkhoff/Bakker 1998, 271–277).

Other stratifications are sometimes used in combination with a genealogical one. A combination of genealogical and areal stratification is used by Tomlin (1986, 24–32). He started from a convenience sample of 1063 languages, which he subsequently reduced to 402 languages to represent the genealogical and areal diversity of the world's languages. For the genealogical stratification, Tomlin refers to Bell's (1978) proposal to represent the number of genera per stock. However, Tomlin actually uses a different method, as his sample represents the number of languages per genus (this means that if a particular genus has, for example, 30 languages out of a total of 6.500 languages in the world, then this genus should ideally be represented in the sample by 30 / 6.500 = 0.46 % of the languages). For the areal stratification, Tomlin used an intuitively established division of the world in 26 areas based on "non-controversial" (Tomlin 1986, 301) areas, limiting the restovers by major continental boundaries (Tomlin 1986, 29). A different approach is a combination of genealogical and cultural stratification as used by Perkins (1992, 129–133). Basically, Perkins included one randomly chosen language per stock, taking care not to take two languages from the same cultural area (for the determination of cultural areas, Perkins refers to an unpublished thesis by Kenny, based on an analysis of cultural traits as proposed by Murdock).

There are various problems with stratified probability samples. The first problem with any stratification is that the resulting sample completely depends on the classification that is followed to obtain the stratification. For example, Rijkhoff and Bakker (1998, 277–292) show that genealogically stratified samples (especially the smaller ones) change drastically depending on the genealogical classification that is used. As genealogical classifications are especially prone to fierce scientific debate, the position taken in this issue will strongly influence any genealogically stratified sample. Second, a stratification is especially effective if the parameter of investigation is known (or expected) to be more homogeneous *within* each stratum than *between* the strata. For example, the 500-odd Bantu languages are strongly homogeneous in having all an SVO basic word order (with one or two exceptions). This homogeneity will substantially raise the number of SVO languages in a genealogically unstratified sample (cf. Dryer 1989, 258). However, genealogical classifications are mostly based on lexicographic and phonological/morphological comparison, which does not necessarily imply a relation to, for example, syntactic properties like word order. So it is not clear beforehand whether a genealogical stratification is of any use for a

typology of a syntactic parameter. Indeed, low-level genealogical strata can show a large variability on syntactic parameters (see, for example, the high diversity of indefinite pronouns within Germanic or Romance, as described in Haspelmath 1997). Such variation within genealogical strata casts doubt on the usefulness of such a stratification. Especially higher levels of genealogical relationship are prone to show a high amount of typological diversity as higher genealogical units are often based on very restricted evidence, leaving much room for variation.

There are two different approaches to deal with these problems. First, Dryer (1989; 1991; 1992) simply ignores all genealogical levels higher than the genus for his stratification. He uses a stratification along genera, but he does not propose any kind of probabilistic representation of genera. He seems to want to include all genera as attested among the world's languages (cf. Dryer 1992, 133–135). He also checks the consistency within each genus by sampling (preferably) more than one language from each genus. Obviously, the disadvantage of this approach is that extremely large samples are needed. In contrast, Perkins (1989) accepts a full genealogical stratification, but he proposes to check its usefulness by statistical calculations. He describes a method that compares the variation within a stratum with the variation between the strata. This method can be used to assess the optimal grain for a stratification. For example, Perkins reanalysed Tomlin's (1986, 301) areal stratification concluding that "the continental grid size displays the maximum effect for word order. Consequently, I infer that continents should be used as the highest level strata for a language sampling frame for basic word order" (Perkins 1989, 309). Note though that this method can only be used *post-hoc*, meaning that only after data have been collected, this method can calculate the optimal stratification.

2.3. Actual vs. possible languages

A general problem for all sample studies is that the best they can do is to represent the *actual* world's languages. However, many investigators would like to use typological samples to make inferences about *possible* human languages. It is conceivable, though, that the actual world's languages are not representative of the possible human languages. For example, having clicks is extremely rare among the world's languages, so typologically speaking there has to be a restricting factor somewhere. However, the question is whether there is an inherent linguistic reason restricting the presence of clicks in the phoneme inventory, or whether there is a completely different rationale for their current distribution. For example, it might just as well have been a coincidence of historical development that those languages with clicks did not spread their characteristics among the world's languages. This possibility implies that the scarcity of a linguistic phenomenon does not necessarily indicate that it is linguistically marked.

The actual and the possible only meet in what Maslova (2000a, 326) calls a 'stationary distribution.' In such a distribution, the net-result of all language change does not influence the frequencies of occurrence of the types; the number of changes between the types is in balance. Maslova (2000a, 315–325; 2000b, 357–361) uses a stochastic Feller-Arley model to investigate the potential effects of changes on the actual frequencies of structural types in a language population. The effects turn out to be most salient in little populations. She concludes that "the current [typological] distributions need not be independent of their initial counterparts. In particular, they may still bear statistically significant traces of those […] events that had happened […] when the language population had been small" (Maslova 2000a, 326). In other words, there is reason to assume that the actual world's languages are not mirorring possible human language.

This position is most forcefully defended by Nichols (1992; 1995; 1996; 1997). She attempts to interpret areal skewing of types as the result of historical processes. To her, typology is the "linguistic counterpart to population biology and population genetics, which analyse variation within and between populations of organisms and use the results to describe evolution" (Nichols 1992, 2). In practice, she compares frequencies of occurrence of linguistic types between various geographical regions (cf. section 4.6.). She is rather eclectic as to which geographic regions she compares, though natural barriers (mountain ranges, large water masses, coastlines) play an important role to delimit the areas. In her more recent works, the areas investigated have become more and more

determined by hypothesised economic/political influences on linguistic distribution.

A different approach to the possible mismatch between the actual and the possible world's languages is to devise a method that controls for this mismatch. Both Perkins (1989) and Dryer (1989) propose such a method by counting only independent cases, i. e. count only those cases of which one is sure that there is no historical connection leading to shared characteristics. Perkins (1989) uses a kind of ANOVA model to analyse the dependency of variation between the languages in the sample according to a particular stratification. He proposes to reduce the sample until there is no significant association any more between the sample and the stratification. Using his method on the 1063-languages sample from Hawkins (1983), Perkins ends up reducing this immense sample to only forty-three genealogically and areally independent cases. This method regularly leads to rather small samples of about 50 languages. Perkins recommends "using around a hundred languages for most linguistic samples to balance the requirements for representativeness and independence in samples. The results from using samples of this size should always be checked, however, to determine if the variables under considerations significantly vary across language groups" (Perkins 1989, 312).

Dryer (1989) proposes an even stronger criterion of independence. He considers only five large continental areas (in later publications there are six areas, see Dryer 1991; 1992) and "the only assumption about independence is that these five areas are independent of each other" (Dryer 1989, 268). Within each area, he counts the number of genera of a particular type, allowing for a genus to be split if its languages are not typologically uniform on a specific parameter (split genera are called 'subgenera', cf. Dryer 1989, 289 n. 4), Any preference should be attested in all areas for it to be interpreted as a linguistic universal (see section 4.5. for a detailed exposition of his method).

2.4. Other approaches to sampling

Maslova (2000a, 328 – 329) describes a completely different method to establish a sample of the world's linguistic variation. She proposes to estimate the transition probabilities between types, i. e. the chances that a language will change its type. These probabilities can be used to compute the stationary distribution of linguistic diversity. Simply put, when a transition from one type to another, say from type T_i to type T_j, is much higher than the opposite transition, from T_j to T_i, then the result will be a proliferation of T_j as compared to T_i. After a long enough period, a stable situation will be reached in which the frequencies of T_i and T_j are proportional to the transition probabilities.

However, there remains the practical problem of estimating transition probabilities. We only have historical information on very few languages – much too few to base any valid estimates on. Circumventing this problem, Maslova proposes two 'apparent time' (cf. Labov 1994, 43 ff.) approaches by using variation to estimate transition probabilities. First, one could use genealogical groups of recent origin and interpret any internal variation to reach an estimate for the transition probabilities. However, it is questionable whether it is really possible to find enough suitable genealogical groups among the world's languages to make a statistically valid estimate. It is also often difficult to infer the direction of change on the basis of variation alone. Maslova's other proposal is to relate the number of 'mixed type' languages in a sample to the number of 'pure type' languages, interpreting languages of a mixed type as intermediate cases in a transition. However, the designation 'mixed type' is highly dependent on theoretical interpretations, as a 'mixed type' might just as well be an unrecognised pure type. Even more problematic, the proportion of mixed type languages in a sample is both a result of transition probability and transition speed. Transition speed is not constant; some languages might stay in an intermediate stage for a long time, while others do not. It seems to be impossible to tease apart probability and speed. And then, even if one succeeds in estimating transition probabilities, it is still possible that these estimates are only valid for the present world's languages. The probabilities of transitions might have been different in the past and might be different in the future.

Compiling a typology of transitions is an interesting approach in itself because it is an attempt to directly investigate the possibilities of language change (cf. Cysouw 2003b, 245 – 294 for an attempt to collect a large set of transitions of the paradigmatic structure of person marking). The results of such an

investigation are *a priori* independent of a synchronic typological survey. Only in a stationary distribution will synchronic and diachronic typologies give compatible results. This implies that all of the above points of criticism also apply to the reverse situation. Not only is it dangerous to deduce synchronic patterns (e. g. universals) for diachronic data (e. g. transition probabilities), but it is also troublesome to infer a typology of change from a purely synchronic typological survey. There is a recurrent attempt in the literature to explain typological patterns with the help of hypothesised universals of language change. For example, Vennemann (1974, 347) proposes to explain exceptions to his typological generalisation of 'natural serialization' by invoking language change. Such conclusions are dangerous, if not downright unwarranted (cf. Mallinson/Blake 1981, 434−435 for detailed criticism to Vennemann). Likewise, Plank/Schellinger (2000) propose to interpret some universals about dual marking diachronically. They are more cautious than Vennemann in their conclusions, but the basic problem remains. Methodologically, a claim about diachronic laws is only possible if the languages in the sample are investigated diachronically. To make a typology of possible changes, one has to investigate a sample of transitions, not a sample of synchronic types.

Finally, a rather different approach towards sampling was pioneered by Plank/Schellinger (1997). They investigated Greenberg's (1963) implicational universals 37 and 45, which (roughly summarised) state that gender in the plural implies gender in the singular. In their study, Plank/Schellinger presuppose this to be true, though they note that there are quite a number of counterexamples, contrary to what is often assumed (see also section 4.3. on the problem of counterexamples). Their attractive approach to investigating the possibilities of human language is to construct a heavily biased convenience sample consisting only of counterexamples to Greenberg's universals. With this collection of 'quirks' they are able to establish deeper insights into the universally valid possibilities of human language. They summarise that "it is hard to know whether the amount of exceptions now on record should cause concern. Encouragingly, it is still with more than chance frequency that gender distinctions prefer the singular over non-singulars […] Nonetheless, when well above 10 % of the languages examined are at odds with what is being predicted […] this would not seem an entirely negligible margin" (Plank/Schellinger 1997, 93). This approach to sampling − collecting examples of cross-linguistically rare phenomena − is also used by Cysouw (forthcoming a, b, c) to investigate typologically unusual patterns of person marking. Other examples are Haspelmath (1994), investigating boundary changes in morphological structure, and Olson/Hajek (2003), investigating the labial flap.

3. Establishing types

3.1. Continuous parameters

Besides the choice of languages, the other basic precondition to establish a typology is to delimit the types. In most current investigations, types are established categorially. In such an approach, there are strict definitions that govern to which type a language belongs. In contrast to this categorial approach, a few authors use continuous parameters. Greenberg (1990 [1954/1960]) was the first to propose non-categorial measurements, in his case to characterise the morphological type of a language. He proposes various indices based on text counts. For example, he defines the degree of synthesis (or 'gross word complexity') of a language as the ratio M/W, where M is the number of morphemes in a particular stretch of text and W is the number of words in the same text. Greenberg's measurements have been refined by Krupa (1965).

Fenk-Oczlon/Fenk (1985; 1993; 1999) use texts counts to test various correlations inspired by Menzerath's Law (cf. art. No. 67) on a cross-linguistic sample of languages. Also Myhill (1992) discusses many different indices based on text counts (cf. art. No. 53). In the same vein, Altmann/Lehfeldt (1973, 71−121) propose many different measurements for all levels of structural analysis, mostly for phonology and morphology, but also a few for syntax. Unfortunately, the measurements proposed by Altmann/Lehfeldt have never been used in a typological survey. In all investigations that use text counts, there is some attempt to control for the type of texts used (for example, often only story-telling monologues are used). Only Fenk-Oczlon/Fenk (1985; 1993; 1999) use translations of a controlled set of utterances for their typological text counts.

In this tradition, a fair amount of work has been done on the question of the explicitness of marking of noun phrases (NPs). Languages differ as to how often full NPs or, converse, zero markers are used to mark arguments. Givón (1983, 17–18) established a continuum of accessibility from zero marked argument, through clitics, dislocated NPs, to full NPs. He hypothesises that, as the contextual identification of an argument becomes more difficult, a construction will be used high on this continuum. He developed various indices to measure the difficulty of contextual identification (Givón 1983, 14–15), like referential distance ('look back'), potential interference ('ambiguity') and persistence ('decay'). For example, referential distance "assesses the gap between the previous occurrence in the discourse of a referent and its current occurrence in the clause […] The gap is […] expressed in the terms of the number of clauses to the left" (Givón 1983, 13). Givón and his co-workers investigated the correlations between the accessibility continuum and the various measures of contextual identification by text counts for a small sample of the world's languages. Later, Myhill (1992, 20–52) extended this approach by using slightly different indices and comparing languages more directly. Bickel (2003) uses a somewhat simpler measurement called 'referential density', which is defined as the ratio of the number of overt argument NPs and the number of available argument positions in a stretch of text. Although he only compares three languages, he also included within-language variability by establishing referential density for various speakers (differencing for age, gender and literacy) of each language. By using ANOVA tests, he found that the between-language variation is bigger than the within-language variation.

A different quantitative approach to type establishment is to combine various categorial parameters into a complex parameter. For example, Nichols (1992, 72–75) uses a ratio of two parameters (viz. counts of head and of dependent marking structures) which both have a range from 0 to 9. Their ratio looks like a continuous parameter, but this is misleading. The original parameters can only take whole numbers as values, so there are actually only 10 different values on each parameter (viz. all integers from 0 to 9). The cross-section of the two parameters result in $10*10 = 100$ different types. Taking the ration of the two parameters reduces the number of possible types to 60 (because some ratios are identical, e. g. $2/8 = 1/4$). Although this is a wealth of types for a typology, strictly speaking it does not qualify as a continuous parameter. The same method of combining categorial typologies into something that looks like a continuous parameter is used by Bakker (1998) to characterise the flexibility of word order in a language (cf. art. No. 59). A major pitfall with the usage of such combined parameters is that they are easily interpreted as indicating linguistic variation on a continuous range. This misinterpretation can lead to erratic explanations (see section 4.1.).

3.2. Semantic maps

One of the central difficulties for the establishment of types is the problem of cross-linguistic comparability. Different languages often have categories and constructions that are alike to each other, yet they are almost never exactly alike. By positing a categorial definition ('a language is of type A if characteristic X is attested, but of type B if X is not attested'), a researcher simply divides the semantic/functional space of variation into two distinct parts. A more detailed typology can be reached by the usage of semantic maps (sometimes called cognitive maps or implicational maps), a method first used by Anderson (1982) to tackle the cross-linguistic variability of marking perfectivity. To establish a semantic map, various (etic) functions of language are distinguished and then for each language in the sample, the (emic) categories or constructions that express those functions are established. A semantic map for such data shows the (etic) functions in a two-dimensional space. Lines connect those functions that can be expressed by the same (emic) category or construction in any language in the sample (see Haspelmath 2003 for a detailed exposition of this method).

A prime example of this approach is Haspelmath's (1997) investigation of indefinite pronouns. He distinguishes nine functions that can be expressed by an indefinite pronoun. Theoretically, with nine functions there are $2^9 - 1 = 511$ combinations possible (the -1 is added because there has to be at least one function covered). However, in a sample of 40 languages, Haspelmath finds 133 indefinite pronouns showing only

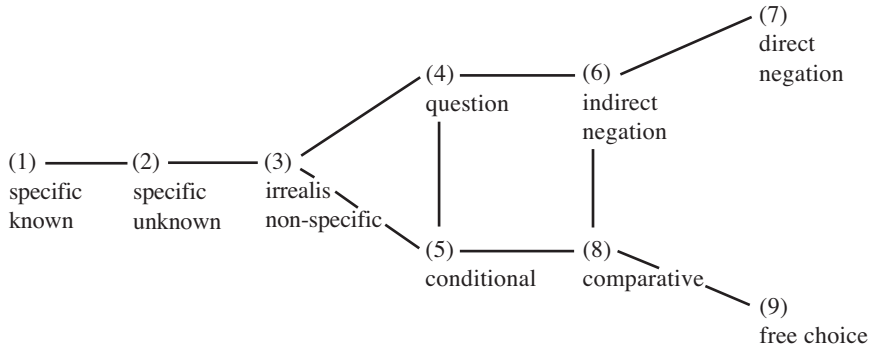

Fig. 40.1: Semantic map for indefinite pronoun functions (reproduced from Haspelmath 1997: 4).

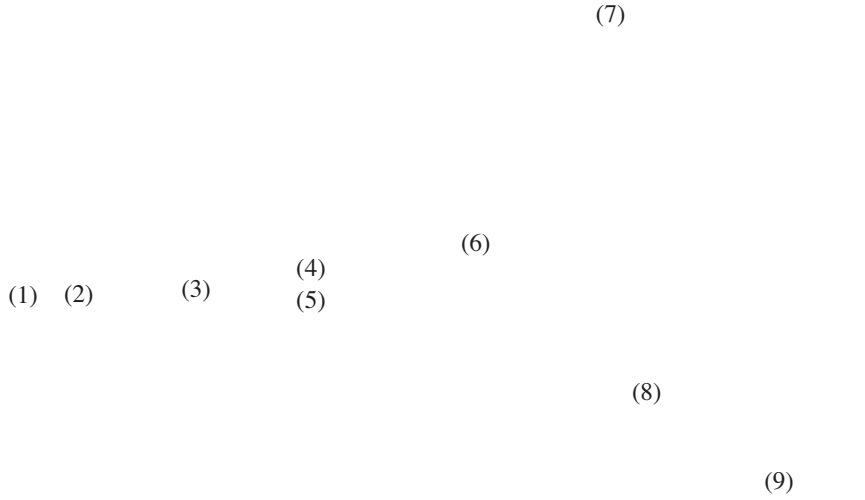

Fig. 40.2: Two-dimensional approximation of the distances between indefinite pronouns functions, based on Haspelmath's (1997: 68–75) data (reproduced from Cysouw 2001: 611).

39 different combinations of functions. To model this apparent restriction of possible combinations, Haspelmath proposes a semantic map, shown here in Figure 40.1. This map restricts the number of possibilities from 511 to 105. With two extra constraints (Haspelmath 1997, 77), this number is further reduced to 82, still about double the number of the 39 cases attested. This model overestimates the variation attested, a recurrent problem with semantic maps.

Another problem with this kind of approach is that a semantic map does not take into account which combinations of features are frequent and which are rare. Cysouw (2001) reanalysed Haspelmath's data incorporating the frequencies of occurrence of the combinations of functions using multidimentional scaling (cf. Croft/Poole 2004). The basic idea is to place the nine (etic) functions in a two-dimensional space in such a way that the distance between any two functions is proportional to their frequency of (emic) co-occurrence. The result of a purely mathematical approximation is shown in Figure 40.2. The similarity to Haspelmath's "inductively established" (1997, 122) map is striking.

3.3. Reducing continuous data

Parameters that use a continuous scale (or distinguish very many types) are often reduced into just a few discrete types. There are two possible reasons to use such a reduction. First, continuous scales might seem somewhat tedious to work with, so a reduction to two or three discrete types can be used to simplify the interpretation of the

data. Second, there is a preference in the typological literature for typologies of just a few types (about 2 to 5) on each parameter. The preference for such parameters is probably based in the widespread belief among linguists that everything essential in linguistic structure will be discrete. Also, many linguists find it conceptually more insightful to interpret a situation of, for example, three different classes of phoneme inventories (small, regular and large) then to work with a range from 11 to 141 phonemes.

At least five different strategies have been used in the literature to reduce a continuous scale of typological variation (or a scale with very many different types) to just a few types of linguistic structure. The basic problem is to decide where to put the cut-off point. The following strategies have been used:

- divide the linear scale into equal parts on the linear scale;
- divide the linear scale into equal sizes of the resulting groups;
- use frequently occurring types on the linear scale as cut-off points;
- use a confidence interval around a central value;
- use similarity as established by a mathematical distance measure.

The first strategy – divide into equal parts – is quite straightforward. For example, Nichols (1992, 98) divides her complexity range with values from 1 to 15 into three equally sized subparts 1–5 ('low'), 6–10 ('moderate') and 11–15 ('high'). Another example is Bakker (1998, 394–405) who divides various parameters with values ranging between 0 to 1 into three equal parts: 0–0.33, 0.34–0.66 and 0.67–1. Another approach with the same result is to round off decimal values. With this method Nichols (1992, 73–74) reduces 35 different classes of head/dependent ratio to 11 classes. The problem with these reductions is that it does not have any intrinsic motivation. It is just a tool to turn continuous data into discrete groups. Fenk-Oczlon/Fenk (1999, 157–158) also use this rounding strategy to reduce a continuous scale describing the average number of syllables per clause. As syllables always come in whole numbers, this reduction might be interpreted as describing the prototypical number of syllables per clause.

The second strategy – divide into equally sized group – at least makes more sense methodologically. For example, Justeson/Stephens (1984, 533) divide the range of phonemic inventories into two parts in such a way that each part consists of the same number of languages in their sample. For example, the cut-off point for the number of vowels in a language is between nine and ten vowels, because 25 out of their 50 languages have nine or less vowels, and the other 25 have ten or more vowels. Maddieson (1984, 10–20) appears to use the same approach, though he does not explicitly state his reasons for establishing his cut-off points. This kind of division still does not mean anything linguistically, but at least the resulting groups are roughly comparable in size, which allows for easier statistical evaluation.

Both the first two strategies are mostly meaningless linguistically. They are purely formal strategies for division, without any reference to content. However, as they are independent of the data, this makes these strategies suitable for all situations. In contrast, the following two strategies can only be used with particular distributions of the data. The third strategy – use frequently occurring types as cut-off points – is for example used by Nichols (1992, 97–98) to simplify her typology of head/dependent ratios, ranging between 0 and 1. She found a few types in this range that were clearly more common than others (but see Cysouw 2002, 78–79 for criticism on this analysis) and she decided to use these types as cut-off points. It remains unclear, though, to which side of the cut-off point these common types should be counted – and this decision strongly influences the frequencies of the resulting types. Yet, in principle a division informed by the actual distribution of the data is linguistically interesting. However, it can only be used if the data show some peaks in their distribution.

The fourth strategy – use a confidence interval around a central value – can likewise only be used with a particular distribution of the data, namely only when the data show a single peak inside the range. Lehfeldt (1975, 284–285; see also Altmann/Lehfeldt 1980, 97–101) divides the phoneme inventory range into parts using a statistical confidence measure around the central peak. In this way, he designates all inventories up to 18 phonemes as 'small' and all inventories from 48 phonemes upwards as 'large'. The same approach is also used by Krupa/Altmann (1966). The problem with this ap-

proach is that it depends on a suitable mathematical model for the data in which confidence intervals can be determined.

Finally, an interesting strategy to establish discrete types from continuous data is used by Altmann/Lehfeldt (1973, 34–48; 1980, 282–293). They use a mathematical similarity measure (combining 10 different continuous parameters) to establish a similarity matrix of their sample of 20 languages. They then organise the languages in a tree in which more similar languages share a node. Their method for organising the languages in the tree is rather outdated, but recent cladistic methods from biology can be used instead (cf. Felsenstein 2004 for a survey). Such a tree of relative similarity can subsequently be used to determine discrete groups of languages by choosing particular branches as establishing a type.

4. Interpreting variation

4.1. One-dimensional skewing

When a sample of the world's languages is established and the languages in the sample are all classified according to the parameter of interest, then the next step is to interpret the frequencies obtained. In almost all typological investigations, it turns out that the various types on a particular parameter are not uniformly distributed. Some types are much more common than others. In the case of continuous parameters (or parameters with very many different types) quantitative models can be of service. For example, the size of phonological inventories varies widely among the world's languages. There is a range from minimally 11 to maximally 141 phonemes with a median between 28 and 29 (Maddieson 1984, 7). Lehfeldt (1975; see also Altmann/Lehfeldt 1980, 87–95) attempts to model the distribution of inventory size using a gamma distribution. This result is criticised by Justeson/Stephens (1984, 538–540; see also Stephens 1984, 651) because there does not seem to be a sensible reason for using a gamma distribution, except that it fits the data rather nicely. Instead, they propose a log-normal distribution, which also fits the data. However, this model also has a motivation: Justeson/Stephens reason that the number of distinctive features used by a language is crucial, not the number of phonemes themselves. They argue that the number of distinctive features used by a language is normally distributed and that "the number of distinctive features exploited in a language is roughly proportional to the logarithm of the number of segments built up from them" (Justeson/Stephens 1984, 539–540). This results in a log-normal distribution for phonemes, which is corroborated by the data.

Building on this distribution of phoneme inventories, Altmann/Lehfeldt (1980, 151–182) show that various characteristics of phoneme distribution are related to the size of the phoneme inventory K of a particular language. For example, the repeat-rate R of a language is defined as:

$$R = \sum_{k=1}^{K} p_k^2 \qquad (2)$$

In this formula, p_k represents the chance of occurrence of a phoneme k in the language. The repeat-rate R describes the mathematical expectation of p_k for all phonemes of a language. If all phonemes were equally frequent in a language (which is counterfactual), it can easily be shown that R would be identical to $1/K$. However, starting from the assumption that the chances for the occurrence of individual phonemes are geometrically distributed, Altmann/Lehfeldt derive that R should be roughly identical to $2/K$. This prediction very nicely describes the actual values of R in a sample of 63 languages (1980, 151–159). Zörnig and Altmann (1983) question the assumption of the geometrical distribution of phoneme frequencies. Assuming a Zipfian distribution (cf. art. No. 16) they get a slightly better fit of the data, though the formula for R is rather complex and loses the intuitive attractiveness of the simpler $2/K$.

There is a major pitfall for the interpretation of skewing of parameters. In some typological studies, the parameters are composites: their values are based on a combination of various empirical measurements (cf. section 3.1.). The interpretation of such composite measures is dangerous, because it is easily forgotten that they are not empirical primitives and consequently the statistically expected values are often not intuitively assessable. For example, the central parameters used by Nichols (1992) are counts of head (H) and dependent (D) marking structures in a language. In her approach to the concept of head and dependent marking, both counts are in principle independent: a structure can be marked on both the head

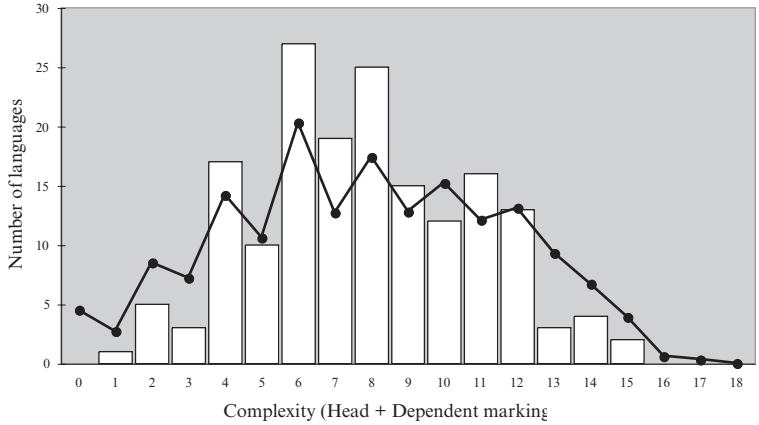

Fig. 40.3: Nichols' complexity data (bars) and the statistically expected values (line), adapted from Cysouw (2002: 77).

and the dependent, only on one of both, or on neither. She uses various composite measures that are based on these counts, such as 'complexity' (defined as D + H) and 'head/dependent ratio' (defined as D / D + H). In the evaluation of these composite measures, Nichols forgets that the statistically expected values are also composites. For example, she finds a "roughly normal distribution" for complexity, which she interprets as "showing that languages avoid the extremes of complexity" (Nichols 1992, 87–88). However, the statistically expected distribution of her value for complexity has about the same form as the attested distribution, as shown here in Figure 40.3 (cf. Cysouw 2002, 77). The same problem is attested with her head/dependent ratio, about which she claims that "it is bimodal, with the greatest peaks at the extremes of exclusive head marking […] and exclusive dependent marking" (Nichols 1992, 72–73). This seems to imply that there is a tendency for a language to be either head or dependent marking. In fact, the distribution described by Nichols is almost purely the result of the transformation D / D + H, not of the underlying data (Cysouw 2002, 78). In these cases, the apparent skewing does not have to be explained or modelled.

4.2. Implicational universals

Although the observation of skewing within a single parameter is already a result of major importance, most typological investigations do not stop there. The ultimate goal of many typological investigations is the establishment of interaction between *a priori* independent parameters. The classical approach to analysing dependency between two discrete typological parameters was introduced by Greenberg (1963) and is concisely set out in Greenberg (1978). The basic tool is the implicational universal A → B, which states that there is an interaction between two parameters A and B in such a way that exactly one of four theoretically possible combination of features, viz. [A+, B−], does not occur among the world's languages. In words, such an implicational universal amounts to saying: for all languages, if a language has characteristic A, then it also has characteristic B. However, if a language does not have characteristic A, then it can either have, or have not, characteristic B (and both these options should occur). This kind of implication is what logicians call 'material implication' and its interpretation is rather different from the intuitive interpretation of the English statement *A implies B*. The intuitive notion makes no claim as to what happens if A is not true. In contrast, the material implication explicitly claims that both B and not-B should be attested when A is not true.

There are two important derivatives of the (material) implicational universal: the bidirectional universal (or logical equivalence) and the implicational hierarchy. The bidirectional universal A ↔ B is a combination of the two mutual implications A → B and B → A. Such a universal claims that two combinations of features do not occur among the world's languages, viz. [A+, B−] and

[A−, B+]. However, Greenberg comments that "statements of this type are hardly ever encountered, perhaps because of their obviousness. They are probably worth more attention in that they involve a very strong relationship, stronger than that of a unidirectional implication" (1978, 52).

An implicational hierarchy consists of "a 'chain' of implicational universals, so that the implicatum of the first universal is the implicans of the second, the implicatum of the second universal is the implicans of the thirds, and so on" (Croft 1990, 96−98). A set of chained universals is shown in (3a). Such a chain is not equivalent (in the mathematical sense) to the nested chain as shown in (3b). The logically accurate way to formulate an implicational hierarchy is shown in (3c). Because this notation is rather cumbersome and uninformative, a hierarchy will normally be summarised by using another symbol instead of the implicational arrow, as for instance shown in (3d). Finally, another equivalent way to depict a hierarchy is shown in (3e). This table shows that five different types of languages have been attested out of $2^4 = 16$ logically possible types. The parameter-settings in this table-like layout intuitively show the hierarchical structure.

(3) a. $A \rightarrow B$
$B \rightarrow C$
$C \rightarrow D$
b. $A \rightarrow (B \rightarrow (C \rightarrow D))$
c. $(A \rightarrow B)$ & $(B \rightarrow C)$ & $(C \rightarrow D)$
d. $A > B > C > D$
e.

	A	B	C	D
type 1:	+	+	+	+
type 2:	−	+	+	+
type 3:	−	−	+	+
type 4:	−	−	−	+
type 5:	−	−	−	−

A central aspect of the implicational approach is that some of the logically possible types do not exist among the world's languages. However, more often than not, it is not as simple as that. Some counterexamples or even whole subregularities are bound to appear among the world's linguistic diversity. One proposal to deal with this is to extend the power of the implicational universal, as put forward by Hawkins (1983). He used concatenations of implications like, for example, in (4) to reach exceptionless statements. By nesting implications, it is possible to include subregularities within a major implicational pattern (see Pericliev 2002 for a refinement of this approach).

(4) If a language has SOV word order, then, if the adjective precedes the noun, the genitive precedes the noun.

Hawkins (1983, 64) uses the notation as shown in (5a), though this is logically equivalent to the expression in (5b). In general, any nested set of implications like (6a) can be reformulated as a single implication with a conjoined implicans, as shown in (6b).

(5) a. $SOV \rightarrow (AN \rightarrow GN)$
b. $(SOV \,\&\, AN) \rightarrow GN$

(6) a. $A \rightarrow (B \rightarrow (C \rightarrow \ldots (Y \rightarrow Z)))$
b. $(A \,\&\, B \,\&\, C \,\&\, \ldots \,\&\, Y) \rightarrow Z$

There are various problems with the usage of such nested implicational universals. First, statements like (5a) seem to imply a relative order of importance among the parameters, though strictly logically this is not the case. This can easily be recognised by considering the equivalent expression (5b). In this expression, the order within the conjunction is not important $(SOV \,\&\, AN) \equiv (AN \,\&\, SOV)$. Thus, in the extended implicational universals the order of all parameters, except for the last, is of no importance (cf. Pericliev 2002, 54 n. 4), The implication $SOV \rightarrow (AN \rightarrow GN)$ is logically identical to $AN \rightarrow (SOV \rightarrow GN)$.

Second, an extended implicational universal is not a very strong statement, contrary to what it might look like at first sight. A statement like (4) seems rather interesting, linking three different characteristics into a meaningful bond. However, logically only one of the eight theoretical distribution of values is actually excluded by this statement (viz. SOV+, AN+, GN−). This weakness becomes even stronger as more implications are nested like in (6a). For each extra level of nesting, the theoretical possibilities double, yet the number of excluded value settings remains the same: no matter how long the concatenation of implications, only one value setting is excluded.

Finally, and most crucially, there is a dangerous empirical pitfall with the usage of nested implications (cf. Dryer 1997, 140−141). The problem occurs if a single implicational universal already presents a strong generalisation. For example, the second nested implicational universal proposed by Hawkins (1983, 65) is shown in (7a). As al-

ready discussed, this is equivalent to (7b). However, if only the second part of this statement, as in (7c), is considered, this universal has only two counterexamples in Hawkins' data, both of them Aztecan languages (but see Campbell *et al.* 1988, 211−212 for a different view of the Aztecan data). Now, it is rather easy to make this implication exceptionless by adding any untrue statement about Aztecan languages as an extra implicans. For example, Aztecan languages do not have tone, so the statement in (7d) does not have any counterexamples anymore. However, being a tone language does not seem to have any sensible typological relation to word order regularities. This example illustrates that by nesting implications one can always get rid of exceptions, but possibly at the cost of adding senseless constraints.

(7) a. VSO → (NA → NG)
b. NA → (VSO → NG)
c. VSO → NG
d. Tone → (VSO → NG)

4.3. The problem of exceptions

From the start, the implicational approach suffered from the problem of exceptions. Greenberg (1963) already qualified some of his implicational universals as holding only "with overwhelmingly greater than chance frequency." For others he did not add this qualification. However, Comrie (1989, 20) rightly commented that "it is virtually impossible in many instances to distinguish empirically between absolute universals and strong tendencies […] either the universal is absolute, or we happen not yet to have discovered the exceptions to it." Dryer (1997) even forcefully argues that statistical universals are better than absolute universals. Many typological studies go into great detail to discount the exceptions to the implicational universals that are encountered. However, empirically this tactic is not legitimate, as it is often just as well possible to cast doubt on the classification of the regular cases (cf. section 2.1.).

Cysouw (2003a) points out two pitfalls for the typologist's disposition towards establishing absolute implicational universals (and their derivatives). The first problem is that non-occurrence of a particular combination of types (the basis of the implicational analysis) does not necessarily mean something. Compare the hypothetical distributions of a 100-language sample as shown in (8). Both distributions show an empty cell in the cross-section of two parameters A and B, which would traditionally be interpreted as qualifying an implicational universal A → B. This works fine for the distribution in (8a), which has one zero and a significant interaction ($p < 0.0001$). However, the distribution in (8b) also has exactly one zero, but does not show any statistical interaction ($p = 0.10$), so the inference is wrong in this case.

(8) a.

	A+	A−
B+	26	48
B−	0	26

b.

	A+	A–
B+	14	72
B−	0	14

I have used Fisher's Exact test here to argue for or against significance of interaction. This test will be used throughout this section. However, this does not mean that this is necessarily the best test to apply to test typological frequencies (cf. section 4.5.). I will report here one-sided exact *p*-values (which is the weakest version of this test). Dryer (2003, 124−126) appears to report one-sided *p*-values for the same or stronger association, and Maslova (2003, 105−106) appears to report two-sided *p*-values for the same or stronger association. The use of either of these depends on the hypothesis that is tested and the resulting values may differ rather strongly for one and the same distribution.

The basic assumption for using Fisher's Exact (or other measures of interaction) is that the parameters are established independently of each other. The test assumes the proportion of A+ to A− and the proportion of B+ to B− is given and calculates on this basis whether there is statistical reason to assume an interaction between the two parameters A and B. It does not say anything about the reasons for any skewed distribution of the parameters in isolation − these remain to be explained. For example, in (8b) no interaction between A and B is attested, so the zero does not have to be explained. However, the distributions of both A and B in isolation are heavily skewed (the plus to minus ratio is 1:6 in both cases).

These skewed distributions still have to be explained.

The second problem with the typologist's focus on implicational universals is that interesting distributions are possibly dismissed because there are no empty cells. Compare the hypothetical distributions of a 100-language sample as shown in (9). Both distributions show ample occurrences of all possible combinations of types, so from the viewpoint of implicational universals there is nothing of interest going on here. However, the distribution in (9a) shows an equally strongly significant interaction as the distribution in (8a) ($p < 0.0001$). For a theory of linguistic structure, this distribution is very interesting. In contrast, the distribution as shown in (9b) does not show any significant interaction ($p = 0.12$).

(9) a.

	A+	A−
B+	35	15
B−	15	35

b.

	A+	A−
B+	26	33
B−	15	26

Cysouw (2003a) concludes from such examples that the presence of any statistical significant interaction is more important than the occurrence of zeros. However, a distribution with both a significant interaction and a zero, like in (8a), remains of special interest. Maslova (2003) describes a useful test to distinguish between different kinds of statistically significant interactions. When there is a significant interaction between two parameters A and B, she proposes to correlate both parameters to a third − derivative − parameter. This new parameter contrasts the cases that are in line with the correlation (i.e. the cases [A+, B+] and [A−, B−], abbreviated below as A = B) with those cases that go against the correlation (i.e. the cases [A+, B−] and [A−, B+], abbreviated below as A ≠ B). There are three different kinds of results when both original parameters are correlated with this new derivative parameter:

- The two extra tests both show a significant interaction. This can be called a *two-sided asymmetrical* dependency;
- Only one of the two tests shows a significant interaction. This can be called a *one-sided asymmetrical* dependency;
- Both tests do not show a significant interaction. This can be called a *symmetrical* dependency.

The three kinds of significant interactions are exemplified in (10) to (12), respectively, In all these examples, the table labelled 'a' shows the original distribution of cases and the tables labelled 'b' show the two additional tests. Both asymmetrical dependencies in (10) and (11) are characterised by one cell that is relatively empty. Asymmetrical dependencies thus resemble the traditional typologist's notion of the implicational universal. However, this is only a superficial characterisation that cannot be reversed (i.e. one cannot deduce from a relatively empty cell that there is an asymmetrical dependency). Maslova (2003) uses the names 'strong unidirectional implication' and 'weak unidirectional implication', respectively, for the two-sided and one-sided asymmetries. However, these labels are misleading in either of the two possible readings. First, there is no difference in the strength of the implication (both are equally strongly significant). Second, 'strong unidirectionality' would most appropriately refer to a more asymmetric situation, which would suggest the one-sided dependency and not the two-sided dependency.

(10) A two-sided asymmetrical dependency

a.

	A+	A−
B+	33	30
B−	4	33

$p < 0.0001$

b.

	A+	A−
A=B	33	33
A≠B	4	30

$p < 0.0001$

	B+	B−
A=B	33	33
A≠B	30	4

$p < 0.0001$

(11) A one-sided asymmetrical dependency

a.

	A+	A−
B+	21	21
B−	4	54

$p < 0.0001$

b.

	A+	A−
A=B	21	54
A≠B	4	21

$p = 0.11$ (n.s.)

	B+	B−
A=B	21	54
A≠B	21	4

$p < 0.0001$

It is tempting to interpret a one-sided asymmetric distribution, like (11a), as showing an influence from B on A, but not the reverse (cf. Maslova 2003, 106). However, it remains to be seen whether the difference between the two-sided and the one-sided asymmetry is linguistically salient. Asymmetric dependencies can be used as a statistically valid replacement of the implicational universal. I propose to use the notation 'A ∼ B' for both asymmetric dependencies, highlighting that there is no direction in the dependency. The symmetrical distribution in (12) also intuitively shows the symmetry in the distribution of frequencies. I propose to use the notation 'A ≈ B' to designate symmetric interactions.

(12) A symmetrical dependency

a.

	A+	A−
B+	33	17
B−	17	33
$p < 0.001$		

b.

	A+	A−
A=B	33	33
A≠B	17	17
$p = 0.17$ (n. s.)		

	B+	B−
A=B	33	33
A≠B	17	17
$p = 0.17$ (n. s.)		

The replacement of the implicational universal with the notion of asymmetrical dependency (as proposed here) poses a problem for the concatenation of such dependencies (as used in implicational hierarchies, see section 4.2.). The problem is that asymmetrical dependencies are not necessarily transitive (in the mathematical sense of the word). Mathematical transitivity states that if a relation holds between the pair (A, B) and the pair (B, C), then it also holds for the pair (A, C). This is true for the material implication: if A → B and B → C, then also A → C. However, this is not necessarily true for the asymmetrical dependency: if A ∼ B and B ∼ C, then A and C do not even have to show a significant interaction! For example, a hypothetical distribution of three parameters A, B and C in a 100-language sample is shown in Table 40.1. This distribution is traditionally interpreted as an implicational hierarchy, based on the material implications (with a few counterexamples). As shown in (13), the interactions A ∼ B and B ∼ C are indeed statistically significant. However, the interaction between A and C is not significant at all.

Cysouw (2003a, 98−99) proposes a different analysis to capture the intuition of a hierarchical distribution in Table 40.1 in a statistically correct way. First, a statistical analysis has to show a significant interaction between the parameters A, B and C. Additionally, the frequencies of occurrence of the three parameters in isolation (as shown in the last column of Table 40.1) have to be significantly different. Indeed, the occurrence of A+ is significantly less then the occurrence of B+, which is in turn significantly less than the occurrence of C+, viz. $21 \ll 50 \ll 79$. The significance of the differences (i. e. that 21 is really significantly smaller than 50, etc.) can be tested, for example, by computing a confidence interval around each frequency. These confidence intervals should not overlap. The combination of a significant three-way interaction and a non-overlapping frequency cline can be interpreted as a hierarchy A > B > C.

A different, and probably better, approach is to use log-linear modelling (e.g.

(13)

	A+	A−
B+	19	31
B−	2	48
$p < 0.0001$		

	B+	B−
C+	48	31
C−	2	19
$p < 0.0001$		

	A+	A−
C+	19	60
C−	2	19
$p = 0.09$ (n. s.)		

Table 40.1: A hypothetical distribution of three parameters A, B and C in a 100-language sample showing an implicational hierarchy.

A	+	−	−	−	+	+	−	+	Total A+ = 21
B	+	+	−	−	−	+	+	−	Total B+ = 50
C	+	+	+	−	−	−	−	+	Total C+ = 79
	18	30	30	18	1	1	1	1	

Justeson/Stephens 1990). However, the details of how to use of such an approach for typology have yet to be worked out. I did some preliminary analyses which indicate that only a very restricted set of results of a log-linear analysis qualify as a hierarchy. Only in case a model with maximally two-way interactions suffices, and these two-way interactions can be lineary ordered, then the data can be said to be modelled by a hierarchy. For example, in Justeson and Stephens' (1990) analysis of word order correlations, they claim to need only two-way interactions for a sufficiently good model. (However, I have not been able to replicate this claim using the data from Hawkins 1983: 288. In my analysis, various three-way interactions were needed to arrive at a good fit.) If this claim from Justeson and Stephens is accepted, then the two-way interactions in their best model can still not be ordered lineary, so there is no hierarchy. The usage of log-linear modelling indicates that hierarchies are rather unusual results, but it also points towards more intricate models that could be fruitfullly used in typology. This is definitively an area that needs more investigation.

4.4. Statistical testing

Although one might think that the collection of typological data will automatically lead researchers to use statistical methods for their evaluation, this is not what has happened. Statistical techniques are used incidentally, but there has not been a general acclaim for the need of such methods. An early example of cross-linguistic cross-parameter statistical testing is found in Krupa/Altmann (1966; see also Altmann/Lehfeldt 1973, 44−48). They investigated Greenberg's (1990 [1954/1960]) morphological parameters in a sample of 20 languages (unfortunately heavily biased towards Indo-European). Correlating the various parameters, they found remarkable dependencies between the parameters (cf. art. No. 58).

As early as 1979, Justeson and Stephens started to look at statistical patterns in word order typology. However, their results were only (partly) published much later (Justeson/Stephens 1990), and even then without any influence on the developments in typology (their work has recently been broughy back into attention by the discussion in Croft 2003: 74−77)

Around 1980, various researchers again started (independently from each other) to use statistical methods to evaluate claims of typological dependency. The first to use basic chi-square testing of Greenbergian-style universals was Isaac Kozinsky in his 1979 Moscow dissertation (as cited in Testelets 2001, 314−316). However, this unpublished and not widely known work did not have any influence on other researchers. Around the same time, Maddieson (1980, 59; 1984, 9) reports on a significant (though weak) cross-linguistic correlation between the number of consonants and the number of vowels in a language. This result is heavily criticised by Justeson and Stephens (1984) using various statistical approaches to argue that there is no such correlation. Also around the same time, Perkins extensively used various correlation coefficients to argue for correlations between linguistic structure and cultural complexity (starting with his unpublished 1980 SUNY Buffalo dissertation, later published as Perkins 1988; 1992). From 1985 onward Fenk-Oczlon/Fenk (1985; 1993; 1999) use statistical methods to show correlations between the size of syllables, words and sentences in a sample of the world's languages. In recent years, statistical tests are used on an off-and-on basis in the typological literature (mostly non-parametric correlation or dependency tests). For example, under Perkins' influence, Bybee also started to use statistical tests to evaluate cross-linguistic frequencies (cf. Bybee *et al.* 1990; 1998).

The most extensive use of correlation statistics in typology to date is found in Nichols (1992). She collected a large typological database to investigate holistic claims made by Klimov (Nichols 1992, 7−12). Nichols performed numerous correlation tests between the various characteristics of the 172 languages in her database. However, in her brave attempt to substantiate all her typological claims with statistical tests, she sometimes forgets to recapitulate the validity of using a statistical test. For example, she correlates two parameters 'complexity' and 'head/dependent proportions' (data repeated here in Table 40.2). Nichols concludes from these data (using a chi-square test) that "head marking favors low complexity and dependent marking favors high complexity. Languages of low complexity show a strong preference to place what little morphology they do have on heads: 21 of

Table 40.2.: Complexity and head/dependent type (adapted from Nichols 1992: 99). Expected values are added between brackets.

Head/Dependent proportions	Complexity levels					
	Low		Moderate		High	
< 0.5 (head marking)	21	(31.1)	50	(35.4)	4	(3.5)
= 0.5	3	(3.5)	5	(4.7)	4	(6.2)
> 0.5 (dependent marking)	10	(13.2)	41	(38.2)	34	(36.3)

Table 40.3: Affix order relative to basic word order (adapted from Siewierska & Bakker 1996: 150, Table 18). Expected values are added between brackets.

	AO		OA		Both		Total
V3	20	(23.3)	18	(12.1)	3	(5.6)	41
V2	20	(14.2)	1	(7.4)	4	(3.4)	25
V1	5	(5.7)	5	(2.9)	0	(1.4)	10
Free	2	(4.0)	2	(2.1)	3	(0.9)	7
Split	3	(2.8)	0	(1.5)	2	(0.7)	5
Total	50		26		12		88

these languages are in the head-marking part of the range, and only ten in the dependent-marking part" (Nichols 1992, 99). However, she forgets that both parameters are based on the same counts of head (H) and dependent (D) constructions. Contrary to Nichols' interpretation, when the statistical expectation is computed from the underlying H and D values (shown in brackets in the table, for computational details see Cysouw 2002, 74−81), it turns out that head marking with low complexity (21 cases) is clearly less common than expected (31.1 cases) and overall there is no significant dependence whatsoever ($\chi^2 = 11.3$, $p = 0.18$).

However dangerous the pitfalls of using statistical analyses, it is still of major importance to check the statistically expected values of any interaction claimed. Statistical significance alone is never enough to qualify an observation as interesting (see section 4.5.). Yet, interpreting numbers without checking chance effects might lead to wrong interpretations. Still, even basic significance tests are not at all standard in typological works. Many authors rely on frequencies and proportions to make their argument. In most cases, the argumentation is not as strongly flawed as the preceding example from Nichols. However, little mistakes can be found regularly. For example, in a typological investigation of verb agreement, Siewierska and Bakker claim that: "in both V3 and V1 languages [but not in V2 languages, MC] AO and OA affixal order is *more or less evenly distributed*" (Siewierska/Bakker 1996, 150 italics added). Judging from their data (repeated here in Table 40.3), AO and OA order in V3 and V1 languages are indeed roughly equally frequent, though surely not evenly distributed. Siewierska and Bakker fail to take into account that there are many more cases of AO (50 cases) than OA (26 cases) in the complete sample. The chance expectation (as added between brackets in the table) reveals that OA in V3 languages is much more frequent than expected (actually 18 cases against expected 12.1) and AO in V3 languages is slightly less frequent than expected (though the deviation form expectation does not seem to be significant here).

4.5. Dryer's approach to significance

Dryer (especially 1989; 2003) opposes the usage of traditional statistical measures (like Fisher's Exact) for typological data because typological samples are often biased: "various examples could be cited from the literature where conclusions are reached, often with levels of statistical significance cited, which can be shown to be artefacts of the nonindependence of the languages in the sample" (Dryer 1989, 265). One of the main reasons for this nonindependence is the existence of strong macro-areal effects in the distribution of linguistic features. These effects, whatever their origin, can distort statistical measures. For example, the languages in Dryer's database show a strongly

significant interaction between the order of adjective and noun, and the order of the negative word and verb (Dryer 2003, 124−126), as shown in (14a). However, the significance appears to be strongly influenced by the languages of North America. Removing this macro-area from the sample results in the disappearance of the significance, as shown in (14b).

(14) a. Whole world

	NegV	VNeg
AdjN	64	12
NAdj	81	43

$p = 0.0025$

b. Whole world, excluding North America

	NegV	VNeg
AdjN	40	12
NAdj	72	41

$p = 0.064$ (n. s.)

Such effects led Dryer to reject non-parametrical statistical tests wholesale and develop a different test for significance, based on the assumption of independence of six macro-areas (see also section 2.3.). This reaction appears to be too strong (cf. Maslova 2003, 102 n. 2), If one finds statistical significance, as in (14a), then it is indeed important to check for areal effects (see section 4.6.), which might disqualify the significance. However, interpreting numbers without checking basic chance effects, as Dryer proposes in his method, might lead to wrong interpretations as well.

To illustrate Dryer's procedure and some possible problems with it, consider the data in Table 40.4, describing the frequencies of the order of verb and object (OV / VO), crossed with the order of the noun and the relative clause (NRel / RelN). This table uses the layout favoured by Dryer, showing a box around the highest frequency of the second parameter (here: NRel / RelN) for each macro-area. To be significant, the same preference should be attested in all six macro-areas.

From these frequencies, there is a clear preference for VO & NRel compared to VO & RelN for all areas (cf. the last two lines in Table 40.4). However, the situation for the OV languages is not as obvious. To argue for a consistent preference among these frequencies, Dryer calculates the proportions of RelN (= RelN / RelN + NRel) for both the OV languages and the VO languages (the results are shown here in Table 40.5). He then compares the values for each macro-area and draws a box around the highest proportion. Now it turns out that all areas show the same preference after all. Finally, Dryer calculates the average of the proportions of all macro-areas (the last column in Table 40.5). By averaging proportions instead of taken the proportion of the average, any overrepresentation of macro-areas is discounted (cf. 'Simpson's paradox').

There are a few problems with this method. First, by splitting up the sample, the number of cases in each macro-area is often too low to reach any significance by itself (even in the extremely large samples that Dryer is using). For example, when the data from Table 40.4 are evaluated using Fisher's Exact (shown in Table 40.6), the complete sample indeed shows a strong dependency between OV and RelN ($p < 0.0001$). How-

Table 40.4: Order of noun and relative clause (reproduced from Dryer 1992: 86).

	Africa	Eurasia	SEAsia&Oc	Aus-NG	NAmer	SAmer	Total
OV&RelN	5	[11]	2	2	3	3	26
OV&NRel	[9]	5	2	[6]	[12]	3	37
VO&RelN	0	0	1	0	0	0	1
VO&NRel	[21]	[8]	[12]	[3]	[11]	[5]	60

Table 40.5: Proportions of genera containing RelN languages (reproduced from Dryer 1992: 87).

	Africa	Eurasia	SEAsia&Oc	Aus-NG	NAmer	SAmer	Average
OV	[.36]	[.69]	[.50]	[.25]	[.20]	[.50]	.42
VO	.00	.00	.08	.00	.00	.00	.01

Table 40.6: Statistical evaluation of the data from Table 4, correlating RelN with OV.

	Africa	Eurasia	SEAsia&Oc	Aus-NG	NAmer	SAmer	Total
$p =$	0.0062	0.0017	0.11	0.51	0.18	0.12	0.0000

Table 40.7: A hypothetical example of an areal breakdown, analysed following Dryer's method.

	Africa	Eurasia	SEAsia&Oc	Aus-NG	NAmer	SAmer	Total
A+, B+	5	1	2	2	3	3	16
A+, B−	9	15	2	6	12	3	47
A−, B+	0	0	1	0	0	1	2
A−, B−	2	2	3	3	2	5	17
B+/A+	.36	.06	.50	.25	.20	.50	.31
B+/A−	.00	.00	.25	.00	.00	.17	.07

ever, only Africa and Eurasia reach significance by themselves ($p = 0.0062$ and $p = 0.0017$, respectively). All other macro-areas do not show any significant interaction. Note that Dryer is counting genera, which will often consist of various languages of the same type, so it might be the case that when counting languages, significance can be reached in other areas as well (yet, he explicitly rejects counting languages, see section 2.3.).

Dryer acknowledges the possible lack of significance of each single area. His method only validates a result when all six areas show the same preference (independently of whether each area in itself reaches any statistical significance or not). He claims that the chance of six independent areas showing the same tendency is low enough to warrant to significant observation. "The logic behind this [method] is that there is only one chance in [32] that all six areas will exhibit a given preference" (Dryer 2003, 110). However, he adds a proviso:

"There are often situations in which one area does not quite satisfy the test […] As a rule of thumb, I adopt the practice of tentatively accepting a pattern as reflecting a real linguistic preference if a type is more common in 5 out of the 6 areas, if the preference for that type is quite strong in those other 5 areas, and if the greater number of genera in the one exceptional area is by a relatively small margin." (Dryer 2003, 112−113).

So, the chances are not really 1 out of 32 (which amounts to $p = 0.031$). Adding six semi-consistent situations (one possibly aberrant case for each of the six areas) results in validation for 7 out of 32 cases (which amounts to $p = 0.22$). This is far from reaching any significance. To counter this objection, Dryer adds the condition that the relative frequencies are important (the preferences in the five consistent areas have to be 'quite strong' and the aberrant case is only exceptional by a 'relatively small margin'). In personal communication, Dryer explains that he has "calculated, under plausible interpretations, that allowing these cases raises p to 0.04 from 0.03, not to 0.22." However, throughout this method, Dryer rejects interpreting the actual numbers: he only looks whether a proportion is higher or lower. But now, only if it is a close call, does he acknowledge that there is a difference between 'strong preferences' and 'small margins'. If such quantitative criteria are allowed, then they should be used throughout, as exemplified by the usage of Fisher's Exact above.

Simply looking for the highest proportion, as Dryer proposes, is a rather crude measure. It might even lead to wrong conclusions, because small differences already count. In the case that all areas consistently show only a small preference, Dryer's method might lead one to conclude that there is an interaction, although statistically speaking there is nothing going on. For example, consider the hypothetical distribution as shown in Table 40.7. In this table, I have changed the distribution from Table 40.4 only slightly (though deliberately into the wrong direction to explain how things could go wrong). The same preference is attested in all six areas, showing a preference for B− in both the A+ and the A− languages. However, the proportions as reported in the lower two lines of Table 40.7 show a clear preference for A+ under con-

dition of B+ (cf. Table 40.5). Such a distribution would lead Dryer's method to the conclusion that there is a significant implicational universal A → B. However, taking the numbers for the total sample, there is no statistically significant interaction at all, when using Fisher's Exact as measure ($p = 0.11$).

Dryer is right in criticizing any ignorant use of statistical measures in linguistic typology. However, his own method − when followed blindfolded − is just as prone to result in errors as a standard statistical test like Fisher's Exact. If one wants to interpret a difference between numbers, whatever their origin, it is always important to make sure that any observed difference is not simply due to chance. It is also of the uttermost importance that typological correlations are investigated as to their areal distribution. What is needed is both statistical significance and areal independence, and these two concepts do not exclude each other (cf. Maslova 2000a, 328; 2003, 102 n. 2). Dryer's method, when handled with care, is a fine approach that attempts to unify both these desideratives into one calculation.

4.6. Areal analysis

Investigating the areal patterns in a typological sample has not attracted much attention in the literature. All approaches, to be described shortly, only test effects in any pre-established areal breakdown of the world's languages. In such a method, the world's languages are first divided into groups based on geographical vicinity, and then these groups are investigated as to internal consistency. However, if no effect is found, there might still be areal consistencies, which happen to be cross-sected by the boundaries of the pre-established areal breakdown. The most basic approach to investigate whether there are any areal patterns at all is to plot a parameter on a world map and look for areal consistencies (cf. Haspelmath *et al.* forthcoming). however, the biggest methodological problem that such a visual analysis of areal patterns faces is to assess the chance probabilities of an areal distribution. It is not at all obvious which kind of areal distribution would be expected based on chance alone. Random distributions in space always appear to show some clustering to the human eye. So, it might very well be the case that the areal patterns attested are to a large extent due to chance.

Perkins (1989) was the first to note that areal patterns can be investigated statistically like any other parameter. He proposed a kind of ANOVA to investigate the effects of a given partition of the world's languages on any observed interaction of parameters. Such an analysis can show an influence of an areal partition on the interaction between linguistic parameters (see also section 2.3.). Dryer (1989) proposed a simpler test (as described in the previous section), based on the principle that any effect should be found in all of the six macro-areas distinguished. A generalisation of Dryer's test for areal effects has been used by Nichols (1992, 187−188). Each worldwide effect should be attested in all areas distinguished − though Nichols allows for an error rate of $p < 0.05$. Assuming that the chances of dominance of a particular feature in an area are binomially distributed, Dryer's test for areal independence becomes a goodness-of-fit test (see Table 40.8 for some selected boundary values). Nichols also reverses this test by looking at the minimal number of departures required for a divergence at $p > 0.10$, which she uses as criterion to show that a particular parameter is areally skewed. The latest approach to testing areal coherence is currently being developed by D. Janssen and B. Bickel (first results were presented in Bickel/Nichols 2003). They use randomization techniques to evaluate whether two areally defined groups of languages are significantly different. This method is especially designed to deal with groups that are strongly different in size (e. g. to compare the languages in one little area to the rest of the world). Such situations makes traditional statistical techniques unreliable.

Table 40.8: Distributions required for significance on Dryer's test (reproduced from Nichols 1992: 188).

No. of areas	Maximum departures allowed for goodness-of-fit at $p < 0.05$	Minimum departures required for divergence at $p > 0.10$
12	2	4
10	1	3
8	1	2
6	0	2
5	0	2
3	not testable	1

Some investigators have used visual approaches to show areal patterns. Van der Auwera (1998a, 1998c) uses an overlay of

various typological isoglosses to make an *isopleth-map*. He uses this method to investigate *Sprachbund*-sized areas, but the method is also suitable for larger areas. The method van der Auwera proposes is to start from a particular language as the standard and encircle those languages that share a particular number of features with the standard. The features themselves need not be identical, only the number of parallels to the standard language is important. The circumference lines thus do not mark identity and are not isoglosses in the strict sense – van der Auwera calls them *isopleths*. Depending on which language is the standard, different maps will appear. Some of these maps will show a geographical contiguous cluster at a high number of features, other maps will not show such clusters. As an example, consider the map in Figure 40.4 (van der Auwera 1998b, 122). This map is based on twelve parameters, all related to phasal adverbials in the languages of Europe. The lines in this figure depict the isopleths surrounding languages that share features on at least ten out of the twelve parameters investigated. Only the strongest clusters are shown, which happen to be the clusters that arise when Danish, German and Bulgarian are chosen as standards. The existence of these clusters is explained by reference to historical contingencies.

An 'inverted isopleth' method is employed by Cysouw (2002, 81–91), reanalysing data from Nichols (1992). This method visualises clusters of similar languages for a chosen area, making it possible to observe differences in the clustering between areas. Some specimens of this visualisation are shown in Figure 40.5. The lines in the pictures encircle linguistic types that are equally common in Nichols' typology, showing clear differences of typological clustering in different areas. A problem with this visualisation is that the data from Nichols are interpreted as continuous parameters, which they are not (cf. section 3.1.).

5. Conclusion

In the last decades, various quantitative methods to capture the world's linguistic di-

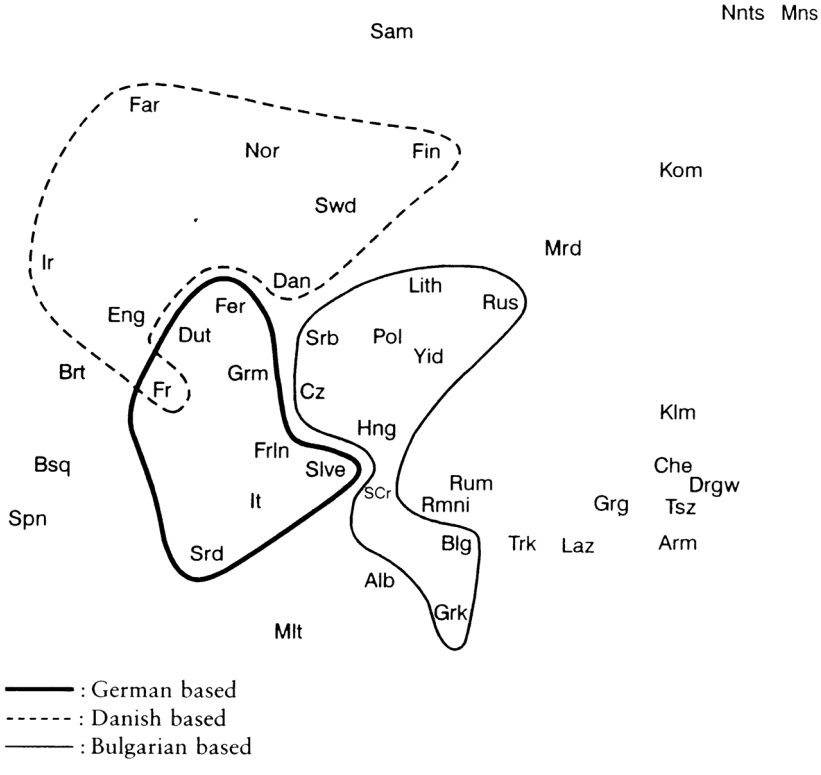

Fig. 40.4: Clustering of 10 or more phasal adverbial parameters, based on German, Danish, and Bulgarian (reproduced from Van der Auwera 1998b: 122).

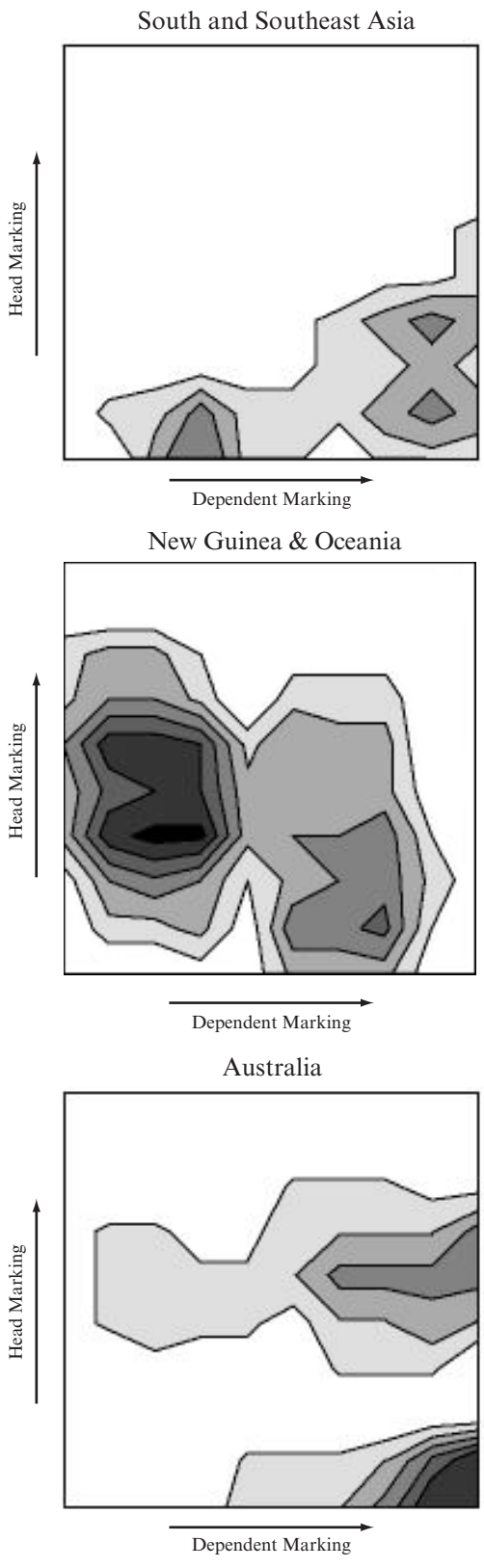

Fig. 40.5.: Clusters of languages with similar head/dependent type in three different areas (reproduced from Cysouw 2002: 86, 88).

versity have been employed in the field of typology. None of them is flawless, but all are sensible to a certain extent. I have attempted to summarise the virtues and pitfalls of these approaches, as used in this flourishing branch of linguistic investigation. The general conclusion is that there is no method that will bring us the holy grail of knowledge just automatically. Investigating the world's languages remains an enterprise in which basic scientific methods like clearly stated hypotheses, consistent argumentation, and careful judgment are more important than fixed methods to be followed blindfolded. The most important work remains very basic: the detailed interpretation of grammatical structures in various languages and the effort to devise parameters that actually allow all those languages to be compared. We are all well advised to follow Plank's scepticism towards fancy statistics: "Nor am I persuaded that doing typology I mostly ought to be doing applied statistics and next to no grammar" (Plank 2003, 138).

6. Literature (a selection)

Altmann, Gabriel & Lehfeldt, Werner (1973), *Allgemeine Sprachtypologie: Prinzipien und Meßverfahren*. München: Fink.

Altmann, Gabriel & Lehfeldt, Werner (1980), *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer.

Anderson, Lloyd B. (1982), The 'perfect' as a universal and as a language-particular category. In: Hopper, Paul J. (ed.). *Tense-Aspect: Between Semantics & Pragmatics*. Amsterdam: Benjamins, 227–264.

Bakker, Dik (1998), Flexibility and consistency in word order patterns in the languages of Europe. In: Siewierska, Anna (ed.). *Constituent Order in the Languages of Europe*. Berlin: Mouton de Gruyter, 383–419.

Bell, Alan (1978), Language samples. In: Greenberg, Joseph H. (ed.). *Universals of Human Language*. (Vol), 1: Method and Theory. Stanford: Stanford University Press, 123–156.

Bickel, Balthasar (2003), Referential density in discourse and syntactic typology. In: *Language* 79 (4): 708–736

Bickel, Balthasar & Nichols, Johanna (2003), Typological enclaves. Paper presented at the 5th

Conference of the Association for Linguistic Typology, Cagliari, September 18, 2003.

Bybee, Joan L. & Chakraborti, Paromita & Jung, Dagmar & Scheibman, Joanne (1998), Prosody and segmental effect: some paths of evolution for word stress. In: *Studies in Language* 22 (2): 267–314.

Bybee, Joan L. & Pagliuca, William & Perkins, Revere D (1990), On the asymmetries in the affixation of grammatical material. In: Croft, William (ed.). *Studies in Typology and Diachrony.* Amsterdam: Benjamins, 1–42.

Campbell, Lyle & Bubenik, Vit & Saxon, Leslie Adele (1988), Word order universals: refinements and clarifications. In: *Canadian Journal of Linguistics* 33 (2): 209–230.

Comrie, Bernard (1989), *Language Universals and Linguistic Typology.* Oxford: Blackwell. (2nd edition).

Croft, William (1990), *Typology and Universals.* Cambridge: Cambridge University Press.

Croft, William (2003), *Typology and Universals.* (2nd edition). Cambridge: Cambridge University Press.

Croft, William & Poole, Keith T. (2004), Inferring universals from grammatical variation: multidimensional scaling for typological analysis. Unpublished Manuscript (available at http://lings.ln.man.ac.uk/Info/staff/WAC/WACpubs.html)

Cysouw, Michael (2001), review of Martin Haspelmath "Indefinite Pronouns". In: *Journal of Linguistics* 37 (3): 99–114.

Cysouw, Michael (2002), Interpreting typological clusters. In: *Linguistic Typology* 6 (1): 69–93.

Cysouw, Michael (2003a), Against implicational universals. In: *Linguistic Typology* 7 (1): 89–101.

Cysouw, Michael (2003b), *The Paradigmatic Structure of Person Marking.* Oxford: Oxford University Press.

Cysouw, Michael (forthcoming-a), Honorific uses of clusivity. In: Filimonova, Elena (ed.). *Clusivity.* Amsterdam: Benjamins.

Cysouw, Michael (forthcoming-b), Syncretisms involving clusivity. In: Filimonova, Elena (ed.). *Clusivity.* Amsterdam: Benjamins.

Cysouw, Michael (forthcoming-c), What it means to be rare: the case of person marking. In: Frajzyngier, Zygmunt & Rood, David S. (eds.). *Linguistic Diversity and Language Theories.* Amsterdam: Benjamins.

Dryer, Matthew S. (1989), Large linguistic areas and language sampling. In. *Studies in Language* 13 (2): 257–292.

Dryer, Matthew S. (1991), SVO languages and the OV : VO typology. *Journal of Linguistics* 27: 443–482.

Dryer, Matthew S. (1992), The Greenbergian word order correlations. In: *Language* 68 (1): 80–138.

Dryer, Matthew S. (1997), Why statistical universals are better than absolute universals. In: *Chicago Linguistic Society* 33 (2): 123–145.

Dryer, Matthew S. (2000), Counting genera vs. counting languages. In: *Linguistic Typology* 4 (3): 334–350.

Dryer, Matthew S. (2003), Significant and non-significant implicational universals. In: *Linguistic Typology* 7 (1): 108–128.

Felsenstein, Joseph (2004), *Inferring Phylogenies.* Sunderland, Massachusetts: Sinauer.

Fenk-Oczlon, Gertraud (1993), Menzerath's law and the constant flow of linguistic information. In: Köhler, Reinhard & Rieger, Burghard B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht: Kluwer, 11–31.

Fenk-Oczlon, Gertraud & Fenk, August (1985), The mean length of propositions is 7 ± 2 syllables – but the position of languages within this range is not accidental. In: d'Ydewalle, G. (ed.). *Cognition, Information Processing, and Motivation.* Amsterdam: North Holland, 355–359.

Fenk-Oczlon, Gertraud & Fenk, August (1999), Cognition, quantitative linguistics, and systemic typology. In: *Linguistic Typology* 3 (2): 151–177.

Givón, Talmy (ed.). (1983), *Topic Continuity in Discourse: A Quantative Cross-language Study.* Amsterdam: Benjamins.

Greenberg, Joseph H. (1963), Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, Joseph H. (ed.). *Universals of Language.* Cambridge, MA: MIT Press, 73–113.

Greenberg, Joseph H. (1978), Typology and cross-linguistic generalizations. In: Greenberg, Joseph H. (ed.). *Universals of Human Language.* Vol. 1: Method & Theory. Stanford, California: Stanford University Press, 33–59.

Greenberg, Joseph H. (1990) [1954/1960], A quantitative approach to the morphological typology of language. In: Denning, Keith & Kemmer, Suzanne (eds.). *On Language: Selected Writings of Joseph H. Greenberg.* Stanford, California: Stanford University Press, 3–25.

Haspelmath, Martin (1994), The growth of affixes in morphological reanalysis. In: Booij, Geert & Van Marle, Jaap (eds.). *Yearbook of Morphology* 1994. Dordrecht: Kluwer, 1–29.

Haspelmath, Martin (1997), *Indefinite Pronouns.* Oxford: Clarendon Press.

Haspelmath, Martin (2003), The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In: Tomasello, Michael (ed.). *The New Psychology of Language: Cognitive and Functional Approaches to Language*

Structure. Vol. 2, Mahwah, New Jersey: Erlbaum, 211–242.

Haspelmath, Martin & Dryer, Mathew & Gil, David & Comrie, Bernard (eds.). *World Atlas of Language Structure.* Oxford: Oxford University Press. forthcoming.

Hawkins, John A (1983), *Word order universals.* New York: Academic Press.

Justeson, John S. & Stephens, Laurence D. (1984), On the relationship between the numbers of vowels and consonants in phonological systems. In: *Linguistics* 22: 531–545.

Justeson, John S. & Stephens, Laurence D. (1990), Explanation for word order universals: a log-linear analysis. In: Bahner, Werner & Schildt, Joachim & Viehweger, Dieter (eds.) Proceedings of the Fourteenth International Congress of Linguists. Vol. 3. Berlin: Akademie Verlag, 2372–2376.

Krupa, Viktor (1965), On quantification of typology. In: *Linguistics* 12: 31–36.

Krupa, Viktor & Altmann, Gabriel (1966), Relations between typological indices. In: *Linguistics* 24: 29–37.

Labov, William (1994), *Principles of Linguistic Change.* Vol. 1: Internal Factors. Oxford: Blackwell.

Lehfeldt, Werner (1975), Die Verteilung der Phonemanzahl in den natürlichen Sprachen. In: *Phonetica* 31: 274–287.

Maddieson, Ian (1980), Phonological generalizations from the UCLA Phonological Segment Inventory Database (UPSID). In: *UCLA Working Papers in Phonetics* 50: 57–68.

Maddieson, Ian (1984), *Patterns of Sound*. Cambridge: Cambridge University Press.

Mallinson, Graham & Blake, Barry J. (1981), *Language Typology.* Amsterdam: North Holland.

Maslova, Elena (2000a), A dynamic approach to the verification of distributional universals. In: *Linguistic Typology* 4 (3): 307–333.

Maslova, Elena (2000b), Stochastic models in typology: obstacle or prerequisite? In: *Linguistic Typology* 4 (3): 357–364.

Maslova, Elena (2003), A case for implicational universals. In: *Linguistic Typology* 7 (1): 101–108.

Myhill, John (1992), *Typological Discourse Analysis.* Oxford: Blackwell.

Nichols, Johanna (1992), *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, Johanna (1995), Diachronically stable structural features. In: Andersen, Henning (ed.). *Historical linguistics 1993*. Amsterdam: Benjamins, 337–355.

Nichols, Johanna (1996), The geography of language origins. In: *Berkeley Linguistic Society* 22: 267–277.

Nichols, Johanna (1997), Modeling ancient population structures and movement in linguistics. In: *Annual Review of Anthropology* 26: 359–384.

Olson, Kenneth S. & Hajek, John (2003), Crosslinguistic insights on the labial flap. In: *Linguistic Typology* 7 (2): 157–186.

Pericliev, Vladimir (2002), Economy in formulating typological generalizations. In: *Linguistic Typology* 6 (1): 49–68.

Perkins, Revere D. (1988), The covariation of culture and grammar. In: Hammond, Michael & Moravcsik, Edith A. & Wirth, Jessica (eds.). *Studies in Syntactic Typology.* Amsterdam: Benjamins, 359–378.

Perkins, Revere D. (1989), Statistical techniques for determining language sample size. In: *Studies in Language* 13 (2): 293–315.

Perkins, Revere D. (1992), *Deixis, Grammar and Culture.* Amsterdam: Benjamins.

Perkins, Revere D. (2001), Sampling procedures and statistical methods. In: Haspelmath, Martin & König, Ekkehard & Oesterreicher, Wulf & Raible, Wolfgang (eds.). *Language Typology and Language Universals: An International Handbook*. Vol. 1. Berlin: De Gruyter, 419–434.

Plank, Frans (2001), Typology by the end of the 18th century. In: Auroux, Sylvain & Koerner, E. F. K. & Niederehe, Hans-Josef & Versteegh, Kees (eds.). *History of the Language Sciences: An International Handbook of the Evolution of the Study of Language from the Beginnings to the Present*. Vol. 2. Berlin: Walter de Gruyter, 1399–1414.

Plank, Frans (2003), There's more than one way to make sense of one-way implications – and sense they need to be made of. In: *Linguistic Typology* 7 (1): 128–139.

Plank, Frans & Schellinger, Wolfgang (1997), The uneven distribution of genders over numbers: Greenberg Nos. 37 and 45. In: *Linguistic Typology* 1 (1): 53–101.

Plank, Frans & Schellinger, Wolfgang (2000), Dual Laws in (no) Time. In: *Sprachtypologie und Universalien Forschung* 53 (1): 46–52.

Rijkhoff, Jan & Bakker, Dik (1998), Language sampling. In: *Linguistic Typology* 2 (3): 263–314.

Rijkhoff, Jan & Bakker, Dik & Hengeveld, Kees & Kahrel, Peter (1993), A method of language sampling. In: *Studies in Language* 17 (1): 169–203.

Siewierska, Anna (1998), On nominal and verbal person marking. In: *Linguistic Typology* 2 (1): 1–56.

Siewierska, Anna & Bakker, Dik (1996), The distribution of subject and object agreement and

word order type. In: *Studies in Language* 20 (1): 115–161.

Stephens, Laurence D. (1984), review of Gabriel Altmann & Werner Lehfeldt: Einführung in die quantitative Phonologie. In: *Language* 60 (3): 650–651.

Testelets, Yakov G. (2001), Russian works on linguistic typology in the 1960–1990s. In: Haspelmath, Martin & König, Ekkehard & Oesterreicher, Wulf & Raible, Wolfgang (eds.). In: *Language Typology and Language Universals: An International Handbook*. Vol. 1. Berlin: De Gruyter, 306–323.

Tomlin, Russell S. (1986), *Basic Word Order: Functional Principles*. London: Croom Helm.

Van der Auwera, Johan (1998a), Conclusion. In: Van der Auwera, Johan (ed.). *Adverbial Constructions in the Languages of Europe*. Berlin: Mouton de Gruyter, 813–836.

Van der Auwera, Johan (1998b), Phasal adverbials in the languages of Europe. In: Van der Auwera, Johan (ed.). *Adverbial Constructions in the Languages of Europe*. Berlin: Mouton de Gruyter, 25–145.

Van der Auwera, Johan (1998c), Revisiting the Balkan and Meso-American linguistic areas. In: *Language Sciences* 20: 259–270.

Vennemann, Theo (1974), Topics, subjects, and word order: from SXV to SVX via TVX. In: Anderson, John M. & Jones, Charles (eds.). *Historical Linguistics*. Vol. 1, Amsterdam: North Holland, 339–376.

Zörnig, Peter & Altmann, G. (1983), The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. In: Köhler, Reinhard & Boy, Joachim (eds.). *Glottometrika* 5, Bochum: Brockmeyer, 205–211.

Acknowledgements

I would like to thank Balthasar Bickel, Bernard Comrie, Matthew Dryer, August Fenk, Reinhard Köhler and Elena Maslova for useful comments on earlier versions of this article. Of course, the opinions expressed in the present version remain completely my own responsibility.

Michael Cysouw, Leipzig/Berlin (Germany)

41. Morphologisch orientierte Typologie

1. Einleitung: morphologische Klassifikation und Typologie

Unter morphologischer Typologie versteht man traditionellerweise den im 19. Jh. einsetzenden Versuch, Sprachen aufgrund *eines* Merkmals ihrer Wortstruktur in verschiedene Klassen einzuteilen (erste Strömung: *klassische morphologische Typologie*). Heute kann morphologische Typologie allgemein so definiert werden, dass sie den morphologischen Bereich der modernen Sprachtypologie meint. Hierbei ist zu unterscheiden zwischen solchen Arbeiten, die sich explizit in die Tradition der klassischen morphologischen Typologie stellen (zweite Strömung: *moderne morphologische Typologie*), und sprachtypologischen Forschungen zur Morphologie bzw. zumeist Morphosyntax (dritte Strömung: (*funktionale*) *morphologische Sprachtypologie*). Als Übergang zwischen den beiden letzteren kann die *strukturalistische Typologie* angesehen werden. Dieser Aufsatz stellt in den folgenden Kapiteln diese verschiedenen Strömungen dar – von der klassischen morphologischen Typologie bis hin zu Arbeiten, die als morphologisch, quantitativ und typologisch charakterisierbar sind. Zunächst erfolgen einige terminologische Bemerkungen zu den Begriffen Klassifikation und Typologie und eine Zuordnung der Ansätze.

Die Begriffe Typologie und Klassifikation werden häufig synonym verwendet, Lehmann (1988, 11 f.) unterscheidet jedoch streng zwischen ihnen. Er versteht unter Klassifikation eine Operation über einer Menge von Gegenständen, durch welche diese in mutuell disjunkte und gemeinsam exhaustive Klassen eingeteilt werden. Dazu sind ein oder mehrere Klassifikationskriterien vonnöten, die gemäß dem Zweck der