Draft, April 15, 2008

A general method for the statistical evaluation of typological distributions

Balthasar Bickel
University of Leipzig

Abstract
The distribution of linguistic structures in the world is the joint product of universal principles, inheritance from ancestor languages, language contact, social structures, and random fluctuation. This paper proposes a method for evaluating the relative significance of each factor — and in particular, of universal principles — via regression modeling: statistical evidence for universal principles is found if the odds for families to have skewed responses (e.g. all or most members have postnominal relative clauses) as opposed to having an opposite response skewing or no skewing at all, is significantly higher for some condition (e.g. VO order) than for another condition, independently of other factors.

Keywords
Language universals, statistical methods, regression modeling, language change, linguistic areas

1. Introduction

Over the past few years, typologists have increasingly addressed problems in the statistical evaluation of proposed universals (e.g. Dryer 2000; Maslova 2000; Cysouw 2003; Janssen et al. 2006; Maddieson 2006; Widmann & Bakker 2006). However, there is still no established methodology in the field, and, somewhat curiously, none of the approaches in current use links up with standard frameworks of statistical analysis that are regularly used in other disciplines. Most surprisingly absent is the family of techniques known as regression modeling, arguably one of the most powerful, and certainly the most successful kind of statistical analysis (e.g. Agresti 2002; for linguistics outside typology, cf. Baayen in press; Johnson in press). In this paper, I propose a way of adapting regression modeling to typological data that solves some of the key problems of statistical typology that have been noted in the past.

The starting point of my proposal is the well-established insight that universals are fundamentally diachronic in nature (Greenberg 1978; Bybee 1988; Hall 1988; Greenberg 1995; Haspelmath 1999; Nichols 2003; Blevins 2004, among many others), and the proposed method is therefore similar to other approaches sharing this starting point, e.g. the approach of Maslova (2000) and Maslova & Nikitina (2007). However, I will argue for a fundamentally different implementation of the insight, one that allows testing hypotheses with multiple factors in competition (a.k.a. 'competing motivations') and also makes less specific assumptions about the nature of diachronic change — crucially, it does not assume constant transition probabilities for typological states.

In the following, I first address the two key challenges to testing universals that have been noted in the past (Section 2): (i) the fact that we

have only ever access to an extremely small and non-random sample of languages from which we would like to extrapolate to distributional skewings in the entire set of languages that our species has ever produced or will ever produce; and (ii) the fact that synchronic distributions are the combined product of multiple diachronic factors, ranging from general inertia/conservativeness to language contact, social factors and universal preferences. In Section 3, I develop a general method for solving these problems by applying multiple regression models to family-level survey data and in Section 4, I discuss technical issues in the implementation of this method. Section 5 illustrates the method by way of a case study on long-standing hypotheses on the distribution of case over word order types (Greenberg 1963; Nichols 1992; Siewierska 1996; Dryer 2002; Hawkins 2004, among others). Section 6 compares the proposed method to alternatives that have been proposed in the literature, and Section 7 summarizes the major components and advantages of the method.

2. Problems of statistical typology

Empirical universals state preferences in the languages of our species that are, by hypothesis, caused by general principles underlying language and language change, ranging from processing principles to principles of communication and principles of self-organization in symbolic systems. An example of an empirical universal is the universal association between verb-object order and postnominal relative clauses, and its hypothesized causes in facilitating processing (Hawkins 2004). Empirical universals differ from absolute universals, which are statements that follow by necessity from the metalanguage ('theoretical framework') employed to analyze languages. An example of an absolute universal is that all languages have distinctive features, or, if one happens to adopt a metalanguage that represents objects as left-hand sisters of verbs, that all languages have an underlying object-verb order.

While absolute universals can be evaluated by applying criteria like descriptive adequacy and coverage, replicability, and logical consistency, empirical universals need statistical evaluation. But any such evaluation is immediately confronted with two key problems:

1. THE INFERENCE PROBLEM: A universal defines preferences for any given language, i.e. for the entire set of languages that our species has ever produced in the past or will ever produce in the future (or at least the set for which one would want to say that it includes human languages the way we know them). The problem is that we cannot take random samples from this set because we have access to only the tiny fraction of languages that happen to be documented right now. If we cannot take random samples, we cannot conduct classical statistical inferences from a sample to the population. How else can we make claims about the entire population?

2. THE DIACHRONY PROBLEM: The distribution of structures that we can observe is the joint product of *structural pressure*[1] ('two languages have both postpositions because they had OV order and then processing became easier with adpositions being postpositional), 'blind' *inheritance* ('two languages have both postpositions because they descend from a language with postpositions, and the postpositions were blindly transmitted, with no regard for anything else'), *language contact* ('two languages have both postpositions because they were spoken by the same people, and people generally prefer a single structure of PPs'), and some degree of *random fluctuation* (cf. Nichols 2003 for a similar decomposition of the relevant factors). How can we separate these different factors, and, most critically for current purposes, how can we distinguish structural pressure from all other factors?

A solution to the Inference Problem can be found if one can solve the Diachrony Problem: if we know that certain diachronic changes are due to structural pressure and nothing else, then we can legitimately extrapolate beyond the currently observable data, because then universals have a time structure that links the past and the future to the observable. If we know, for example, that the observed distribution of postpositions is driven by preferred pathways of diachronic change (and not, say, the contingencies of language contact), then we can legitimately expect that these preferences were the same in the past; if they weren't, they wouldn't have led to the distributions that we observe. And it is reasonable to expect that universals of change will be the same in the future, *ceteris paribus*.

Therefore, the key problem to be solved is the Diachrony Problem: how can we distinguish universal pressure on change from all other diachronic processes? It helps to decompose this problem into three more specific and better solvable sub-problems:

1. THE AREALITY PROBLEM: how can we identify language contact effects?
2. THE RESIDUALS PROBLEM: how can we identify random fluctuation and fluctuation caused by unknown factors?
3. THE INHERITANCE PROBLEM: how can we identify blind inheritance effects?

In the following, I first address the Areality and the Residuals Problem, and then the Inheritance Problem.

---

[1] Other appropriate terms are 'selection', 'functional pressure', 'preferred pathways of change', 'linguistic principles'. I am not concerned here where exactly any such pressure is grounded: perhaps it is hard-wired in the brain, perhaps it results from communicative and social principles. Also, I am not concerned with the question whether structural pressure affects typological distributions by selecting preferred outcomes of random change or by pre-defining pathways of change. For various positions on these issues, see in particular, Haspelmath 1999; Kirby 1999; Croft 2000; Blevins 2004;

2.1 The Areality and Residuals Problems

The Areality and Residuals Problems are statistically relatively trivial as soon as we reformulate linguistic universals as proper statistical hypothesis. The standard way of doing this in other disciplines is by means of multiple regression models, and there is no reason not to do this in typology as well. Multiple regression models allow the identification of the effect of areality as opposed to structural pressure, and at the same time an identification of that part of the distribution that cannot be explained by a hypothesized factor because it is due to random fluctuation and unknown factors.

Multiple regression has a generalized form that is applicable to any kind of variable, including the kind of binary and multinomial variables that are common in typology. The first step in transforming universals of the classical form '$p \rightarrow q$' (e.g. 'VO word order implies an increased likelihood of postnominal relative clauses') into a regression model is to think about q in terms of E(q), i.e. the mean value one expects it to have, given certain values of p (the hypothesized predictor, or series of predictors). With continuous responses, E(q) can (mostly) be directly linked to the predictors, but because of their specific distributional properties, the expected values of categorical and count variables are usually first transformed by what is called a link function. The most commonly used link function for binary categorical responses is the natural logarithm of the odds of the expected response, i.e. $\log(\frac{\pi(q = \text{A})}{1-\pi(q \neq \text{A})})$, where the expected response is the mean (proportion) of q to have value A (e.g. 'VO oder'). This is called 'logistic regression' and also extends to multinomial categorical responses which can be decomposed into sets of binary ones. For count (frequency) responses, one usually takes the logarithm of the expected mean count, log(E(q)), a transformation leading to what is called 'loglinear analysis'.[2] Representing the link function by g, and assuming that one expects no error, the generalized linear model is:

(1) $\quad g(E(q)) = \alpha + \beta_1 \cdot p_1 + \beta_2 \cdot p_2 + \ldots + \beta_k \cdot p_k$

In (1), α (known as the intercept) represents the baseline estimate of q if all $\beta_{1\ldots k}=0$, while the coefficients $\beta_1 \ldots \beta_k$ estimate the relative effect of a series of predictor variables $p_1 \ldots p_k$, including their interactions (and possibly some nonlinear transformation of some predictors or their interactions). What is left unaccounted for by $p_1 \ldots p_k$ is then due to random fluctuation and unknown predictors. This amount can be estimated by a conventional statistic of predictive strength (e.g. R^2 and its variants).

---

[2] In the following I mostly use logistic regression because it offers an easier interpretation for universals and area affects, and because most hypotheses on record involve only few and mostly binary variables, but nothing that follows depends on this choice. Since categorical variables define cell counts in contingency tables, loglinear analysis is another option. It was used once in typology by Justeson & Stephens (1990), but these authors did not attempt to solve the areality problem with this (but see Perkins 2001 for cursory suggestions). Note that all I say here about regression models is textbook wisdom; for good expositions targeted at a linguistics (though not typology) audience, see Baayen (in press) and Johnson (in press).

The predictor variables *p* can include various factors suspected to compete in how they influence the response *q*. These can be various structural variables, e.g. word order or the number of relevant distinctive features in phonology — or just as well some social factor like population size or marriage systems hypothesized to affect the distribution of linguistic structures. Crucially for current purposes, one of the predictor variables can be a linguistic area. To illustrate, (2) is the multiple regression version of the universal 'if a language has VO instead of OV order, it is far more likely to have postnominal than prenominal relative clauses', factoring in the possibly confounding effect of some area distinction, e.g. languages in Eurasia vs. languages outside Eurasia. Here, E(*q*) are the odds for having postnominal relative clauses ('NRel') as opposed to prenominal relative clauses ('RelN'), and *g* is the logarithmic function:

$$(2)\quad \log\left(\frac{\pi(\text{NRel})}{\pi(\text{RelN})}\right) = \alpha + \beta_1 \cdot \text{VO} + \beta_2 \cdot \text{AREAS} + \beta_3 \cdot \text{VO} \cdot \text{AREAS}$$

Once a universal is formulated in this way, the problem is to estimate whether $\beta_1 \ldots \beta_k$ are different from zero to a statistically significant degree — a problem that we cannot solve until we have also addressed the Inheritance Problem. Before proceeding to this discussion, a few more clarifications about (1) and (2) are in order.

First, categorical predictors in regression models are often binary, e.g. 'VO vs. OV order' or 'Eurasian vs. other languages', and are mathematically entered into models with values 1 vs. 0, arbitrarily chosing one category as the baseline (0) against which the effect of the other (1) is compared. Multinomial predictors with *k* levels can be reformulated as *k*-1 binary parameters, again choosing one level as the baseline: for example, if we wish to model the impact of four macroareas like Africa, Eurasia, Australasia and Americas, this can be formulated as binary parameters like [Eurasia vs. Africa], [Australasia vs. Africa] and [Americas vs Africa], with Africa as the arbitrary baseline. The impact of the macro-area factor is then represented by a vector of individual parameter coefficients (e.g. $\beta_{2,1}$[Eurasia vs. Africa], $\beta_{2,2}$[Australasia vs. Africa], $\beta_{2,3}$[Americas vs Africa]), instead of one single coefficient.

Second, the product of predictors, here VO · AREAS, is their interaction and its coefficient (β_3) represents the differences in effect of one predictor across the levels of the other predictor. This can be interpreted either as the difference in effects of VO in Eurasia vs. outside Eurasia (since β_1VO $+\beta_3$VO·AREAS = $(\beta_1+\beta_3$AREAS)·VO), or of Eurasia among VO order vs. other orders (since β_2AREAS$+\beta_3$VO·AREAS = $(\beta_2+\beta_3$VO)·AREAS). The two options can be examined by a follow-up analysis ('factorial analysis') of each equation separately (see Section 5 for an example). With multinomial predictors, interactions are again represented by vectors of binary parameters, one for each difference in effects of one predictor across the levels of another predictor. For example, with four macro-areas and one binary word order factor, this defines (4 – 1) · (2 – 1) interactions, interpretable for example as [VO in the Americas vs. in Africa], [VO in Eurasia vs. in Africa], and [VO in Australasia vs. in Africa]. If it turns out that the resulting interaction coefficients $\beta_{3,1} \ldots \beta_{3,3}$ are simultaneously different from zero, VO order will not have a uniform impact on the odds for postnominal relative clauses, and one will reject the hypothesis of a principle

that holds universally, i.e. independent of the location of languages and their contact histories.

While the Areality Problem is statistically trivial because it can be reformulated as a standard regression problem, the Areality Problem is of course linguistically anything but trivial — indeed, it is arguably one of the most pressing research questions in modern typology. The crucial challenge is to identify the kind of area that can plausibly affect the distributions of interest. This challenge is not specific to research on universals, and it is orthogonal to the problem of how we can statistically evaluate empirical universals. However, one issue is worth noting for current purposes:

Linguistic areas are traditionally defined by sets of structural isoglosses. Yet the conclusiveness of these isoglosses rests on the assumption that they are not universally correlated (e.g. Masica 2001). This leads to circularity: we need to know universals before we can test area hypotheses, and we need to know areas before we can test universal hypotheses. A response to this is proposed by Bickel & Nichols's (2006) 'Predictive Areality Theory'. In this approach, areal hypotheses are grounded outside linguistic structure, in population history. For example, we know that Eurasia has seen repeated spreads of objects, ideas, and languages, often carried by male-dominated military and commercial expansions (e.g. Nichols 1998; Nasidze et al. 2003; Chaubey et al. 2006; Rootsi et al. 2007; and the archeology of the Silk Road). It is plausible that this has led to a large number of language contact events, and this can be formulated as a testable hypothesis of Eurasia as an area which can be directly entered into a regression model.

Instead of actual areas, one can of course also model the impact of specific contact scenarios, e.g. language shift vs. borrowing (Thomason & Kaufman 1988), or different socio-geographical profiles like spread zones vs. accretion zones (Nichols 1997). The model itself is neutral as to what factors are considered.

2.2 The Inheritance Problem

Given the way areality can be modeled through multiple regression, one is tempted to try and model inheritance in the same way: if there is faithful inheritance within families, then membership in families will be a good predictor of current distributions.

In some research designs, family membership can indeed be successfully built into a regression model. In a study of the mean size of phonological word domains, Bickel et al. (in press) model the impact of blind inheritance, represented as family membership, along with the impacts of areality and a structural factor:

(3) $E(\mathrm{c}) = \alpha + \beta_1\mathrm{STRESS} + \beta_2\mathrm{AREAS} + \beta_3\mathrm{FAMILIES}$

Here, c is an approximately continuous variable representing the ratio of morphemes included in a phonological domain divided by the possible maximum in a given language (e.g. c=1 means that the phonological domain spans the entire grammatical word, c=.5 that it only includes half of it; 'c' is mnemonic for 'coherence'). The factor STRESS classifies phonologi-

cal patterns as to whether they are defined by stress vs. something else. The factor AREAS is defined by two binary parameters Europe vs. South Asia and Southeast Asia vs. South Asia. The factor FAMILIES is defined by two binary parameters Indo-European vs. Austroasiatic and Sino-Tibetan vs. Austroasiatic. Because the sample is not a random sample, we cannot apply classical sampling theory to test factors for statistical significance. But we can subject (3) to Monte-Carlo (i.e. randomized) permutation testing, in order to estimate the probabilities of finding the observed coefficients and, for the multinomial factors, observed vectors of coefficients, under the null hypothesis of independence (Janssen et al. 2006; Bickel et al. in press).

This test procedure revealed a significant main effect of family and a significant main effect for STRESS, but no effect for area and no effect for any interaction (which are therefore left out from the formula above). Such a finding entails that the within-family variance is smaller than the between-family variance, and a plausible interpretation of this is that languages of the same family have fairly faithfully inherited their *c*-values, with only little fluctuation.

This approach allows one to factor out the relative impact of inheritance and structural pressure on the development of the current distribution of *c*: the development must have been affected by both inheritance of a fairly uniform *c*-value per family, and at the same time by structural pressure to develop or retain *c*-values that systematically differ between stress-related and other sound patterns. Crucially, the two factors do not interact, and the hypothesized pressure therefore holds independently of family membership.

This way of assessing the relative impact of inheritance and structural pressure has a severe limitation though: it only works if one limits the dataset to a carefully selected sample with a handful of families, each containing a comparable number of languages or relevant structures. There is no way of knowing whether some suspected structural pressure is in fact limited to the few families studied and may perhaps have no effect in other families. If we find the effect in many different families we can have some confidence that it reflects a genuine universal — at least to the degree that there is no plausible alternative interpretation for why STRESS has the same effect across unrelated families and independent of areas.

However, simply adding more families to a model like (3) is not a solution because *k* families need *k-1* binary parameters for regression modeling. The result would be an uninformative model in which the number of parameters approaches the number of datapoints. (In fact, for all single-member families, the number of parameters is identical to the number of datapoints.) To avoid this problem, we need an entirely different approach.

As many typologists have noted, and as I have tacitly assumed in the preceding discussion, universals are best understood as systematic pressures on how languages change over time to form new languages (e.g., Greenberg 1978; Bybee 1988; Hall 1988; Greenberg 1995; Haspelmath 1999; Maslova 2000; Nichols 2003; Blevins 2004). The core idea is that, if there is a universal principle at work, dispreferred distributions will be removed during these processes of change, e.g. after sufficient time, most VO language with prenominal relative clauses will change into languages with postnominal relative clauses.

In order to transform this idea into a statistical modeling procedure, one can rely on the notion of a family as defined for the Comparative

Method, i.e. as sets of diachronic innovations. Each of these innovations can be affected by universal principles, either by favoring a certain innovation (e.g. from VO to OV order) or by mitigating against it. If many innovations in many families are affected in this way by universal principles, this will lead to what I call here 'family skewing': there will be more families that have innovated structure in such a way as to end up skewed in the way predicted by the universal and less families that end up not being skewed (i.e. internally diverse) in this way or being kewed in the opposite way (cf. Nichols 2003; Maslova & Nikitina 2007): within each family, languages will either develop from a dispreferred state into the preferred one, or, if they already are in the preferred state, they will keep that state. In the case of VP order and relative clauses, this would mean that families with VO order will end up skewed towards postnominal relative clauses; whereas families with OV order will be diverse or skewed in either direction. (Families with both orders pose a special problem that will be discussed in Section 4.1 below.)

If no universal is at work, there can be either of two outcomes: (i) Structures may be inherited faithfully from the parent to the daughter languages, regardless of any conditions — e.g. languages may keep prenominal relative clauses regardless of whether the parent language had VO or OV order. If structures are inherited in this way, there is no innovation, and families end up skewed in whatever way the proto-language happened to be skewed. To the degree that this is the case, families will be equally skewed in any direction, i.e. we expect as many VO families skewed towards prenominal as towards postnominal relative clauses (which evidently is not the case, since only one family — Sinitic — is known to have VO order and to be skewed towards prenominal relative clauses). (ii) Another possible outcome in the absence of a universal principle is that there is some innovation in the relevant structure when a parent language splits up, but this innovation shows no particular preference: given a VO parent language, daughter languages would then just as likely develop prenominal as they would develop postnominal relative clauses. The choice may be random or a result of unknown (perhaps areal) factors. In either case, the family would end up diverse (as is the case with relative clause positions in Formosan, apparently as a result of varying degrees of contact with Sinitic).

In summary, if one finds that nearly all families in a survey show the same skewing under specific conditions (e.g. nearly all VO families are skewed towards prenominal relative clauses), this can be interpreted as evidence for universal pressure. If this is not what one finds, but families are skewed in diverse ways even under the same conditions (e.g. some VO families are skewed towards prenominal, some towards postnominal relative clauses), or if they are mostly diverse, then there is no evidence for universal pressure. I call this mode of inferencing ‘the Skewed Family Method’.

To what extent is this inference method valid? Suppose we find the same skewing in virtually all families worldwide — e.g. almost all VO families are skewed towards postnominal relative clauses (as is indeed the case) —, and we interpret this finding not as a reflex of universal pressure, but instead as due to blind inheritance, i.e. in each family, it just happens that the proto-language had VO order and postnominal relative clauses, and this was simply kept by all or most daughter languages. It follows that the current skewing can then only have arisen if the proto-languages had

a similar worldwide skewing as what we find now. But then, how did the generation of proto-languages arise? If again by blind inheritance, the proto-proto-generation would have again had shown a similar worldwide skewing; if it hadn't, there must have been universal pressure to change the distribution in a systematic way. Now, it is logically possible that the proto-proto-generation, indeed that all earlier generations in the set of what we call human languages, had similar distribution as the current one. If that was the case, then the overall probability of random, non-directed change must be exceedingly small. As a result of this low probability of change, we then expect not to be able to observe changes within the relatively short time interval covered by the Comparative Method and almost all reconstructible families will show absolute uniformity in the variable of interest (e.g. relative clause position). Yet in many cases we do find that families evidence changes with regard to structural variables (i.e. one or more languages deviating from the proto-language), and the more we find evidence for change within families, the less is it likely that a worldwide skewing trend across families results from blind inheritance over many generations.

It is instructive to estimate the probabilities of random change p_r that would need to be assumed if a systematic worldwide skewing is interpreted as the reflex of blind inheritance so that cases of change can still be detected. There is a lower and an upper boundary condition on p_r:

1. The probability p_r must be high enough so that we can expect to observe changes in the known set of reconstructed families. In large databases, the size of this set can go up to about 130 families; often it is less than 50. (For example, applying the AUTOTYP taxonomy of reconstructible families to Dryer's (2005) large word order database, reveals 131 highest-level taxa).

2. The probability p_r must be low enough so that an initial skewing is still detectable after a number of random changes that approximates the age of human language. This number is unknown, but it has a plausible minimum of 100, on the account that human language is at least 100Ky old (probably much older in fact) and that structural change (of, say, word order) happens no more often than every 1Ky or so.

The lower boundary of p_r (as per Condition 1) can be determined by assessing how many cases of change we can expect to find in 130 families by chance alone: if p_r=.01, for example, we can expect to find at most 3 cases, or with p_r=.10, at most 18 cases. In a set of 50 families, p_r=.01 leads one to expect at most 2 cases, p_r = .10 at most 9 cases. For each of these p_r–value and sample sizes, finding any more cases would be unexpected, i.e. significant under a binomial test. Thus, if we find more than 9 cases of change in 50 families, we can infer that p_r cannot be smaller than .10. Two real-world examples: in Dryer's database on relative clause position (Dryer 2005a) 11 out of 51 families show evidence of change (i.e. at least one family member differs from all others). For this to be expected, p_r must be at least .13. In a combined dataset on the relative order of A (transitive agent) and O (object) (AUTOTYP and Dryer 2005b), there are 130 families with more than one member. Of these, 55 show evidence of

change. For this to be observable by chance, p_r must be at least .35. This suggest that for most variables, a reasonable lower boundary is $p_r \geq .10$

For estimating the upper bounds of p_r (as per Condition 1 above), I performed computer simulations. Each simulation starts with a dataset of the same magnitude as the largest available databases (about 1300 languages) and assumes an initial skewing that is statistically detectable by a χ^2-test, e.g. a 30% vs. 70% distribution of values. This dataset is then sent through 100 'generations', where at each generation, a random proportion of languages equal or smaller than p_r is changed (thus acknowledging the fact that the rate of language change is not constant over time). For example, given p_r=.01, one generation may change the maximum of .01*1300=130 languages, but the next generation may affect only 20 (or perhaps none) of them.[3] Changes from one to another value are equiprobable in the simulation, because any difference in probabilities would presuppose the force of some universal principle, i.e. the exact opposite of what the simulation aims to model. The simulation program then determines how likely it is that the initial skewing is still detectable by a χ^2-test after 100 generations. This likelihood is computed by counting how often the skewing was detectable in a large sample of simulations (N=1000).

Running these simulation sets with various values for p_r and various initial distributions shows that at p_r=.01, the initial skewing is almost always still detectable after 100 generations. But at p_r-levels closer to what one usually observes in available databases, e.g. p_r=.10, the likelihood that an initial skewing is still detectable after 100 generations falls below the conventional .05 threshhold of random success, and this holds regardless of how strong the initial skewing was (ranging in the simulations from 0%:100% to 40%:60%). This demonstrates that interpreting a worldwide uniform skewing across families as the result of blind inheritance requires assumed probabilities of language change that are by order of magnitude below what one normally observes. This excludes blind inheritance as a realistic avenue of explanation. To the extent that worldwide uniform skewing across families is statistically significant, we can also exclude random fluctuation as an explanation. Such family skewing patterns are therefore best explained as the result of structural pressure, i.e. genuine universals of language. What is still missing in this, however, is a control for areal confounding factors. How this control can be built into the method is the topic of the following.

3. A general model of universals

The preceding discussion suggests that distributional skewings in families reflect signals of structural pressure. This can be directly formulated as a statistical hypothesis: structural pressure is statistically evidenced to the degree that families are skewed in the proposed direction under a hypothesized structural condition (e.g. skewed towards postnominal relative clauses only under the VO word order condition). Possible competition from language contact, social structures and other patterns can be directly built into the hypothesis if we formalize it as a regression equation of the

[3] The program was written in R (R Development Core Team 2008) and relies on R's built-in pseudo-random number generator.

following kind (where *L* represents a linguistic structural factor and *A* a language contact area or some other confounding factor):

$$(4)\quad \log\left(\frac{\pi(\text{proposed skewing})}{\pi(\text{opposite}\,|\,\text{diverse})}\right) = \alpha + \beta_1 \cdot \mathrm{L} + \beta_2 \cdot \mathrm{A} + \beta_3 \cdot \mathrm{L} \cdot \mathrm{A}$$

Here, datapoints are not languages but entire families (with more than one member each), classified as to whether or not the distribution of the response variable of interest (e.g. relative clause position) is skewed conditional on *L* and/or *A*. For a hypothesized universal to get statistical support, (4) must have a coefficient β_1 (or, with multinomial factors, a vector of parameter coefficients $\beta_{1,i}...\beta_{1,k}$) that is significantly different from zero and must not have an interaction coefficient β_3 (or vector of interaction coefficients $\beta_{3,i}...\beta_{3,k}$) that is significantly different from zero, i.e. we expect *L* to skew families independently of *A* (across different areas, or social structures, or whatever is modeled by *A*). In Section 4, I propose an algorithm for measuring the skewing across families, and I discuss statistical problems associated with finding and testing the coefficients in (4). Before going into these more technical issues, however, I wish to clarify the nature of hypothesis formulation that (4) is meant to capture.

The model in (4) is suitable for both unidirectional ('$p \rightarrow q$') and bidirectional ('$p \leftrightarrow q$') hypotheses. These two types of universals differ in the expectations about the odds ratio: For a unidirectional hypothesis, it is sufficient that the odds for the proposed skewing is higher for one level of the predictor than for the other (as directly reflected by a positive value of β_1, hence a large odds ratio $\theta=e^{\beta_1}$), e.g. higher for VO than for OV families. Crucially, the hypothesis is compatible with a scenario in which the odds under one of the predictor levels (e.g. OV) is 1:1 (which seems to be the case with relative clause positions: the odds for RelN and NRel skewings seem to be roughly the same for OV families). This is different for bidirectional universals. Consider the universal: 'if a family is consistently VO rather than OV, this increases the odds for a skewing towards prepositions; and, if a family is consistently OV rather than VO, this increases the odds for a skewing towards postpositions'. Here, we expect that the odds for a preposition vs. postposition skewing do not approach 1:1 under either level of the predictor; instead, we expect that the odds for preposition as opposed to postposition skewing are many:1 under VO and 1:many under OV.

The model in (4) also subsumes univariate universals as a special case. Univariate universals, e.g. Greenberg's Universal Nr. 1 predicting a universal preference for Agent-before-Object order (Greenberg 1963), contain no linguistic structural predictor but only a baseline frequency distribution α and some areal predictors whose possibly confounding influence we wish to test. A univariate universal is statistically supported if the best-fitting model only includes α. Whether α is skewed itself can then be assessed by a χ^2–test against what is expected under the null hypothesis (e.g. a 1:1 distribution).

If there is statistical evidence for a hypothesis modeled as in (4), we have good reasons to assume that there is universal structural pressure at work, and we can even estimate the time interval in which the universal exerts its pressure on language change: this time interval is always the same as the interval captured by the genealogical taxonomy used. If this

is Dryer's (1989) genus level, then the universal must have exerted its pressure within some 2,000 years; if the model is applied to a taxonomy of stocks in Nichols' (1997) sense, i.e. the deepest reconstructible taxa, then a found universal must have exerted its pressure over a time depth in the magnitude of stock ages, i.e. up to about 6,000 years. In other words, if we find systematic skewings of stocks, we can conclude that a universal has skewed a sufficient number of families within less than about 6,000 years.

However, there could also be universal structural pressure that has slower effects than this, i.e. the pressure might skew diachronic change only over the time course of many more generations of languages than what the Comparative Method allows one to reconstruct. In such a case, (4) will fail to show a significant effect of a the structural effect *L* that is tested in the model. Instead, the distribution of structures within families will be determined by one of the following events: (i) Within the time-frame of the assumed taxonomy, daughter languages blindly inherit whatever happens to characterize the proto-language, regardless of any structural conditions; this will approximate a 1:1 odds for the proposed vs. the opposite skewing, leaving almost no room for diverse families. An example that comes close to this is the distribution of gender (Nichols 2003 and the data in Corbett 2005): families are likely to be skewed towards having gender or towards not having gender; freely 'mixed' families are relatively rare. (ii) Daughter languages diversify in response to unknown factors and/or by random fluctuation; this will approximate a 1:1 odds for the proposed skewing vs. diversity within families, leaving almost no room for families with the opposite skewing. (iii) There is a mix of both unknown factors and faithful inheritance, yielding roughly uniform frequencies of families with the proposed, those with the opposite and those without any skewing.

If what we observe is close to (i), we are confronted with exactly the situation that prompted Dryer (1989; 2000) to develop a principled method of genealogical sampling, i.e. one that controls for the multiplication of features (variable values) that can happen to families as a result of inheritance within the time depth of the taxonomy. In such a case, we need to reduce our sample in such a way that each stock that is skewed as a result of inheritance is represented only once. An algorithm achieving precisely this is developed in Bickel (in press), elaborating on Dryer's (1989) proposal. If we are willing to assume that the inheritance pattern found among non-singleton families can be generalized to the prehistory of isolates, isolates can also be included in the dataset (as is usually done). The resulting sample can then be evaluated again by standard regression modeling, but now with sample languages rather than families as datapoints.

If such a model has coefficients significantly different from zero, and there is no evidence for an interaction with areas, this is a possible pointer to a deep time universal that exerts pressure on diachronic change within larger intervals than what is covered by the assumed genealogical

taxonomy.[4] However, in this case, we can have only much less confidence in the finding, because the stock representatives and isolates in the dataset may happen to be the sole survivors of what were unskewed (diverse) stocks before, or, worse, deviating survivors of stocks skewed in the opposite direction. I will return to the issue of how the proposed method in (4) compares to genealogical sampling in Section 6.

The other scenarios mentioned above (a skewing in the opposite way than what the model predicts, or mixed results) do not open avenues of research for deep time universals. Rather, they suggest that the tested model does not suit the data. Under Scenario (iii) (mixed results), one is well-advised to entertain entirely different models. But Scenario (ii) suggests that the model is on the right track, and only makes predictions in the wrong way: there appears a systematic dispreference for families to be skewed in the way coded as 'opposite'. An example of this is what one observes with the distribution of accusative vs. ergative alignment in case systems (cf. Nichols 1993, 2003, Maslova & Nikitina 2007). In general, the odds for families to be skewed towards accusative alignment is roughly equal to the odds for families to be diverse or to be skewed towards ergative alignment. Thus, if one takes 'accusative alignment' as the 'proposed' value in a model of the kind given in (4), there won't be a significant effect. However, the odds for families to be skewed towards ergative alignment are extremely low, and at any rate much lower than the odds to be skewed towards the opposite (accusative alignment) or to be diverse. This suggests a universal principle disfavoring ergative alignment. (These findings are tentatively corroborated by a survey of AUTOTYP data on 25 families, but further research is needed, on databases covering more families.)

4. Implementation of the method

In order to develop a statistical method for testing the equation in (4), we need two ingredients: (i) an algorithm that estimates which families are skewed in which direction, (ii) tools for assessing the probability of nonzero coefficients without making random-sampling assumptions. I take up these issues in turn.

4.1 Estimating family skewing

In some cases, distributional skewings within families can be determined in a straightforward way. The skewing may be absolute, e.g. all members may have prenominal relative clauses; or all member may have the same

[4] This by and large resolves the debate between Maslova (2000) and Dryer (2000): on the one hand, there is good justification for Dryer's concern that blind inheritance can lead to artificially skewed distributions if a sample contains large families, but this concern is only relevant if inheritance is blind to universal pressures within the time depth of families (i.e. if families are skewed in diverse ways). On the other hand, there is good justification for Maslova's concern that Dryer's sampling strategy throws away critical data for detecting universals, but this concern is only relevant if universals exert their pressure within the time depth of known families.

degree of synthesis. When there is diversity, skewing can be determined by a statistical criterion. For categorical responses, a suitable criterion is a permutation test based on χ^2-deviations from what is expected under the null hypothesis (e.g., equal probability, or probabilities predefined by the definition of the variables involved); in the case of continuous response variables, a possible criterion is to test how often the observed variance is below the variance obtainable in bootstrap samples (samples with replacements) from the full range of possible values.

Determining family skewing becomes more difficult when families are not uniform with regard to the predictors in the regression model, as when, for example, Sino-Tibetan has both VO and OV orders and straddles two linguistic areas of interest (Southeast Asia and South Asia). How can family skewing be determined in such cases? The Skewed Family Method can detect structural pressure at any given time depth because it is neutral as to the taxonomy on which it is applied. If the method detects a statistical signal from structural pressure within shallow families, this suggests that the relevant pressure has effects at a relatively quick pace of diachronic development. If the method detects a signal only at higher-level taxa, this suggests that the pressure affects distributions at a slower pace. Either case is evidence for structural pressure as a universal principle. Indeed, any taxonomic level is just as good a probe for the method as another. Therefore, when a family is split across predictors levels at the highest taxonomic level, it is methodologically legitimate to assess skewing at a lower level, which may not be split. This is so in the Sino-Tibetan example with regard to word order: there are two major branches that are uniformly VO (Karenic and Sinitic), but all other major branches are uniformly OV. The same logic applies to splits by areas: some major branches are in one area, some in another area.

However, given the often sketchy knowledge that is available on subgrouping it is often impossible to find plausible subgroups; or, even though the taxonomy may be well established, subgroups may be diverse with regard to some predictor of interest. In these cases, I propose to posit pseudo-groups based on the difference in predictor values, e.g. a VO pseudo-group vs. an OV pseudo-group. Importantly, these pseudo-groups are posited solely for the purposes of testing whether differences in the predictor have an effect on the distribution of some response variable within each group. They clearly are not evidence for real subgroups. However, since some change must have split the family, it is a legitimate isogloss for testing purposes: the key question is only whether the isogloss is associated with different responses to such an extent that the pseudo-groups are now skewed.[5]

Another problem arises when predictors are continuous, e.g. when taking degree of synthesis, or number of consonants as a predictor for some structural distribution. For this, the only available solution is to slice the predictor into broader categories (e.g. low vs. mid vs. high synthesis degree) and then determine response skewing within each genealogical unit that receives a uniform category assignment.

---

[5] An algorithm that determines skewing with families, with any number of predictors, is available as an *R* function 'families()' in www.uni-leipzig.de/~autotyp/gsample3.r. I thank Taras Zakharko for implementing the algorithm.

4.2 Estimating and testing regression coefficients

As argued in Janssen et al. (2006) and noted in Section 2.2 above, a fundamental problem for any statistical method in typology is that datasets are not random samples from an underlying population. Instead of classical random-sample inference, the only possible type of inference that can be applied in such cases involves permutation methods: the significance of an observed distribution is determined by comparison to random permutations of the observed data itself. In other words, the null hypothesis is that the observed distribution is just as likely as the distribution under any re-shuffling of values in the data.[6]

Permutation tests can be applied to any statistic. In the case of regression models, one method is to randomly permutate the response, i.e. the relative frequencies of families with the proposed skewings as opposed to those with the opposite skewing and those with no skewing. For the observed dataset and for each permutation of it, one then computes the likelihood ratios *LR* of nested models,[7] in which the best fitting coefficients (i.e. the values of α and $\beta_{i...k}$ in 4 that best predict the data) are estimated via standard Maximum Likelihood estimation (e.g., Agresti 2002). The *LR* statistic (also known as 'deviance' or 'G^2') measures the difference in data fit between two nested models and is defined as the difference between $2\log\Lambda_1$ and $-2\log\Lambda_2$, where Λ_1 and Λ_2 are the maximum likelihoods of the two models.[8] A common case of interest would be the likelihood ratios between a model including an interaction between a structural and an areal factor and a model without such an interaction. The statistical significance of the *LR* of the models — in our example, the difference between the more complex model including the interaction (with $\beta_3 \neq 0$ in 4) and the less complex model excluding the interaction (with $\beta_3 = 0$ in 4) — is then given by the number of cases in which the *LR* statistic in the permutated datasets is at least as high as the *LR* statistic obtained in the observed dataset. If that is the case in more than, say, 5% of a large number of permutated datasets (e.g. 10,000), the *LR* statistic is not significant. In our example, higher *LR* statistics will arise with those permutated datasets that are better fitted by a model with interactions than by one without.

If the *LR* between two models of the observed dataset is often matched or surpassed by the *LR* between the same two models of random permutations, this suggests that the likelihood difference could have

---

[6] Alternative terms focus on various aspects of the same method: 'conditional inference' focuses on the fact that all inference is conditional on the observed dataset, 'exact test' focuses on the fact that p-values are determined in comparison to all possible alternative datasets ('approximatively exact', if the comparison involves only a random subset of these alternatives), 're-sampling' focuses on the fact that many samples are drawn from the same dataset, and 'randomization' on the fact that permutations are random. See Everitt & Hothorn 2006; Good 2006; Manly 2007, among others.

[7] Models are nested iff the less complex model is a subset of the more complex model and contains all terms presupposed by the interaction terms in the more complex model.

[8] For sparse datasets with many predictors, maximum likelihood estimation may not work well and should be replaced by conditional likelihood estimation, see Agresti (2002: Chapter 6.7), Forster et al. 2003, and Zamar et al. 2007 for solutions. A convenience function for performing permutation tests based on likelihood ratios is available for R in www.uni-leipzig.de/~autotyp/rnd.lr.test.r. The function is compatible with any kind of regression model and any kind of variables.

arisen by chance alone and that the two models fit equally well (or equally badly!). Applying Occham's razor, the less complex model is then preferred; in our example, there is then no evidence for an interaction between area and structure, i.e. β_3 is not significantly different from zero in (4).

To determine the significance of each individual factor of a regression model and each interaction in it, one can perform such a test of significance for the *LR* statistic comparing a model with the term of interest and one without. Testing of successively smaller models then leads to the most parsimonious model compatible with the data. Once one finds this model, one will also want to assess its over-all fit by comparing it to what is known as the 'saturated' model, i.e. one which contains as many predictors as it has data and therefore fits perfectly and trivially (e.g. each language predicts its own response). If our most parsimonious model fits as well as the saturated model (so that the *LR* between the two models is not significant under a permutation test), it is a good description of the data. 'Good' here can of course only be understood relative to the hypothesis under investigation. An entirely different set of predictors, i.e. a different theory, may always be a superior description!

5. A case study

Many typologists have hypothesized that verb-final order favors what I call here 'A≠O marking', i.e. case or adposition marking distinguishing A ('subjects', transitive agent-like arguments) from O ('objects') (e.g. Greenberg 1963: Universal Nr. 41; Nichols 1992; Siewierska 1996; Dryer 2002; Rijkhoff 2002; also cf. Konstanz Universals Archive Nr. 447). Hawkins (2004) discusses explanations for this in terms of increased efficiency of incremental processing when arguments are overtly distinguished before the verb is processed.

However, typologists have also noticed that the worldwide distribution of both case/adposition marking and of word order is heavily influenced by language contact, resulting in strong areal patterns (Dryer 1989; Siewierska 1996; Dryer 2000, 2005b; Bickel & Nichols 2006, in press, among others). For example, Eurasia is known to favor case whereas Africa is known to disfavor it. Southeast Asia and Europe are known to favor VO order while the rest of Eurasia is known to favor OV order.

The critical question then is whether the distribution of A≠O marking is driven by word order (specifically, the difference between verb-final vs. other orders), independently of both areas and blind inheritance within families. Assuming the method developed above, the issue can be formulated as a regression model (VF = 'verb-final vs. non-verb-final', A= 'areas')

$$(5)\quad \log\left(\frac{\pi(\text{skewed towards A} \neq \text{O})}{\pi(\text{skewed towards A} = \text{O} \mid \text{diverse})}\right) = \alpha + \beta_1 \cdot \text{VF} + \beta_2 \cdot \text{A} + \beta_3 \cdot \text{VF} \cdot \text{A}$$

The hypothesis then is that β_1 is significantly different from zero —perhaps along with β_2 — but that β_3 is not significantly different from zero, i.e. that an interaction between word order and area does not improve the fit of

the model and can therefore be neglected. If this is so, there is evidence that the factor VF affects language change in such a way that families tend to be skewed towards distinguishing A and O by case or adposition marking.

5.1 Data and Coding

The data for testing (5) come from merging the datasets from AUTOTYP (Bickel & Nichols 1996ff) and the *World Atlas of Language Structures* (specifically, Comrie 2005; Dryer 2005b), classified into linguistic areas at various levels of resolutions and into a genealogical taxonomy contained in AUTOTYP (cf. above).[9] Merging seems legitimate since the databases converge in the coding of those languages covered by both. For word order (final vs. non-final order, excluding variable and free orders), the coding converged in all 207 such cases; for argument marking (A=O vs. A≠O), the coding converged in all but one of 100 such cases.[10] The resulting set covers 330 languages, with 51 families containing more than one member. This is not much, but will do for illustrating the method.

Given what is known from the literature about the geography of case and word order, it is not self-evident what level of areal resolution is plausible. In response to this, I tested the impact of *A* at three levels of resolution: I first examined a breakdown of the world in 24 traditionally-sized linguistic areas (e.g. Southeast Asia, Europe, California) and deviating remnant regions (e.g. Caucasus, North Australia) (Test 1). These are the kinds of areas which have often been noted to affect the distribution of word order. Second, I tested a 4-way breakdown of the world into 'macrocontinents' in the spirit of Dryer (1989) and Nichols (1992) (Test 2). Third, since the distribution of case is particularly affected by the Eurasian macro-area (Jakobson 1931, Bickel & Nichols, in press, and Section 2.1 above), I examined a two-way distinctions between languages in Eurasia vs. others (Test 3). (Following Bickel & Nichols 2003, I excluded the Caucasus and the Himalayas from the Eurasian spread area.) Maps 1-3 identify these geographical breakdowns.

INSERT MAP 1 ABOUT HERE

Map 1: Areas assumed for testing purposes in Test 1 (A = Alaska-Oregon, B = Andean, C = Basin and Plains, D = California, E = Caucasus-Mesopotamia, F = Eastern North America, G = Ethiopian Plateau, H = Europe, I = Indic, J = Inner Asia, K = Interior New Guinea, L = Mesoamerica, M = N Africa, N = N Australia, O = North Coast Asia, P = North Coast New Guinea, Q = North Savannah, R = Northeastern South America, S = Oceania, T = Southern

[9] The data and all codings are available at www.uni-leipzig.de/~autotyp.

[10] The one mismatch concerns the African language Fur (ISO 639-3: fvr), where accusative case distinct from the nominative (the so-called 'compound accusative') is limited to some verbs (Beaton 1968). The merged dataset represents Fur as a language with A=O marking, but this decision has no impact on the results.

Africa, U = Southern Australia, V = Southern New Guinea, W = Southeastern South America, X = Southeast Asia)

INSERT MAP 2 ABOUT HERE

Map 2: Macrocontinents in Test 2 (stars = Africa, squares= Americas, dots = Eurasia, triangles = New Guinea and Australia)

INSERT MAP 3 ABOUT HERE

Map 3: Eurasia in Test 3 (black dots; without the Caucasus and the Himalayas)

5.2 Results

I first tested a model with the 24-way areal breakdown (Map 1). Determining family skewings necessitates pseudo-groups in 59% of cases in order to derive families with uniform predictor values, $N = 94$. In total, 78% are on the highest taxonomic level, the others on lower levels. The skewing distribution is plotted in Figure 1.

INSERT FIGURE 1 ABOUT HERE

Figure 1: Distribution of family skewing per area (Test 1, same labels as in Map 1). The width of each area-labeling box is proportional to the sample size of the area. Within each area, the bars to the left display non-verb-final, the bars to the right verb-final order. The width of the bars is proportional to the number of families under each condition (zero is represented by a line with a round circle). Within each bar, the black part represents families skewed towards A≠O; the grey part represents families that are skewed in the opposite way or diverse (i.e. unskewed) families.

As shown by Figure 1, some interactions of area and word order are undefined because only a single word order is found in the area. Data from these areas need to be removed before it is reasonable to fit a model with interactions. This results in 14 instead of 23 degrees of freedom for testing the significance of the interaction coefficients.[11] There is no evidence for an interaction term (LR=13.98, df=14, p=.89), but there are significant main effects for both the word order factor (LR=17.83, df=1, p<.001) and

[11] An additional problem is that the relative large number of parameter coefficients (N = (2-1)+(15-1)+(2-1)·(15-1) = 29) and partial collinearity between them can lead, and with the given data, does lead to computational problems in Maximum Likelihood Estimation. In order to avoid this, I followed standard recommendations (cf. e.g. Harrell 2001 or Baayen, in press), and built a penalizing factor into the model fitting algorithm before performing tests on the obtained likelihoods. The best-matching factor was empirically determined to be 3.

the area factor (LR=52.92, df=23, p<.003). The best-fitting model therefore includes both these factors but without interactions; comparing this additive model to a saturated one suggests a good over-all fit (LR=55.16, df=69, p=.99). The odds ratios of the word order factor is θ=35.47, i.e. under this model, verb-final families are about 35 times more likely to be skewed towards A≠O marking than other families.

Results are similar for the 4-way macrocontinent breakdown (Test 2, Map 2). Here, determining family skewing necessitates pseudo-groups in 31%, N=77. In total, 62% are on the highest taxonomic levels and 23% on the next-to-highest level. Figure 2 shows the skewing in families across the four macrocontinents.

INSERT FIGURE 2 ABOUT HERE

Figure 2: Distribution of family skewing per macrocontinent (Test 2, same plotting conventions as in Figure 1)

There is no evidence for an interaction term (LR=2.72, df=3, p=.53), but there is a significant main effects for word order (LR=13.20, df=1, p<.001) and a marginal effect for the macrocontinents (LR=7.32, df=3, p=.07). The overall-fit of an additive model is good (LR=82.45, df=72, p=.98). The odds ratio of the word order factor in this model is θ=6.93, i.e. under this model, verb-final families are almost 7 times more likely to develop a skewing towards A≠O marking than other families.

Figure 2 suggests that the word order effect is strongest in Eurasia. This observation can be further examined by building the difference between Eurasia and the rest of the world into the model, but now defining Eurasia as a spread zone, without the Caucasus and the Himalayas (Test 3, Map 3). For such a model, determining family skewings requires 33% pseudo-groups, N=79. In total, 58% are on the highest taxonomic level, 41% on the next lower level. Figure 3 displays the observed distribution.

INSERT FIGURE 3 ABOUT HERE

Figure 3: Distribution of family skewing per macrocontinent (Test 3, same plotting conventions as in Figure 1)

The difference in the strength of the word order effect is confirmed by a borderline significant interaction (LR=4.15, df=1, p=.054). Factorial analysis of the word order effect inside and outside Eurasia suggests that the skewing has the same direction and is significant in both (Fisher Exact test, Eurasia p=.002, Other p=.003). This suggests that an additive model might fit just as well as one with an interaction. Such a model fits the data reasonably well (LR=88.04, df=76, p=.99).

5.3 Summary

In all three tests, areal factors make a significant contribution to the skewing of families towards A≠O marking. However, while the strength of this effect varies, it does not interact with the hypothesized word order effect in such a way that it would reverse it. In other words, the word order effect always has the same direction. The effect is statistically significant in all models, and this lends evidence to the hypothesis that the development or maintainance of A≠O marking within families is indeed correlated with verb-final order. This points to universal structural pressure in the way families have developed over time.

6. Discussion

How does the proposed method compare to alternatives that are available in the literature? There are two dimensions in which my proposal differs from previous ones: (i) it employs regression modeling in order to control for areal and other factors; (ii) it controls for inheritance effects by determining distributional skewings within families. I take these issues up in turn.

The classical alternative to regression modeling is to separately examine individual areas (Dryer 1989). This is the same procedure that is standardly applied in factorial analysis when there is evidence for an interaction (as was the case in Test 3 above). A well-known problem of this procedure, however, is that the individual sub-samples may be too small for revealing any association between variables (also cf. Cysouw 2005). For example, if instead of modeling regressions, I had performed four separate Fisher Exact tests on each macrocontinent in Test 2, the results would have suggested that it is only in Eurasia that word order has a significant effect on A≠O marking (p=.002); in all other areas, the effect is not (Africa, New Guinea-Australia) or borderline (Americas, p=.06) statistically significant. A natural conclusion from this would be that the word order effect is not universal since in some areas it can be predicted from the margin totals (i.e. the total proportion of verb-final and of A≠O marking families). However, the results of these individual tests are a side-effect of the considerably different sample sizes, as visually represented in Figure 2 by the length of the area-denoting boxes under each plot. When area is controlled for in regression modeling, word order has a significant impact on A≠O marking, with an appreciable odds ratio of around 7.

The classical alternative to examining family skewing is genealogically balanced sampling, where the data are reduced in such a way that each genealogical unit is represented by the sole or predominant value of some variable of interest (Dryer 1989, 2000; Bickel in press). The problem of this method is that it assumes that all skewings or uniformities within families are the result of blind inheritance from their respective proto-languages. As argued in Section 3, this is only the case if (a) skewings go in different directions, independently of structural factors, and (b) together outrank family-internal diversity. If the skewings depend on structural factors or do not outrank family-internal diversity, the distributions within families are the best data we have for assessing the significance of these factors. Reducing the sample then means to throw away all critical data. In return, if

a response variable is skewed within families independently of structural or areal factors, the variable is extremely stable, and genealogically balanced sampling is an excellent method to control for this stability.

There is one situation, however where genealogically balanced sampling is the only option available: the method proposed here requires that each family be represented by more than one language. This entails that isolates can never enter the analysis. For some factors of interest — e.g. A≠O marking and word order — this is not a problem. But it quickly turns into a problem when the variables of interest happen to be best represented in isolates. For such research questions, genealogically-balanced sampling is the only possibility. Also, genealogically-balanced sampling obviously has a very practical advantage because it can also be performed before collecting data, thereby reducing the workload in creating a database (see Bickel in press for some discussion of this kind of 'a-priori' sampling).

In summary, genealogically-balanced sampling still deserves an important place in quantitative typology (*pace* Maslova 2000). When it is applied, it is important, however, to note the limitations of the method. The most severe limitation is that the method offers no 'dynamic' interpretation of universals: while the Skewed Family method proposed here gives direct evidence for universals as principles of diachronic change, the genealogical sampling method only takes synchronic snapshots, as it were. From these, we cannot infer principles that drive the development of typological distributions over time. They may be indicative of such developments, but there is no guarantee.

7. Conclusions

This paper makes a new proposal on how to assess whether an empirical universal holds independently of other factors known to influence the development of typological distributions: the impact of blind inheritance of whatever type the parent languages had (e.g. the position of relative clauses) can be assessed by determining the directions in which families are skewed, possibly under other factors of interest (e.g. VP order or areas). The impact of areas can be factored out through regression modeling, i.e. a statistical method that is standard in other disciples, including close neighbors like psycholinguistics and sociolinguistics.[12]

The statistical problem that the data are not random samples can be solved by applying permutation methods and conducting conditional inference limited to the data. Since the data represent diachronic change probabilities (skewing towards certain features under given conditions), statistical significance of a factor directly attests to its diachronic relevance. If a factor does indeed play a universal role in diachrony, it is plausible to assume that it projects into the past and the future. Under this assumption it is in turn possible to infer principles that are truly universal, i.e. independ-

[12] For historical reasons, regression modeling is known as VARBRUL in sociolinguistics. In psycholingustics, regression modeling is typically restricted to continuous response and categorical predictor variables and the term ANOVA is used as shortcut for models with just this kind of design. See Johnson (in press) for discussion.

ent of time and space, and this overcomes the inference limitations of permutation methods.

Acknowledgements

Previous versions of this were first presented in 2006 at the Leipzig Spring School on Linguistic Diversity and the Australian Linguistics Institute and then in various lectures in Leipzig in the Winter of 2007 and a guest lecture in Helsinki in March 2008. I thank these audiences for helpful questions, and I am in particular grateful to Stefan Lang and Dirk Janssen for discussions of regression and permutation problems, respectively. Many thanks are also due to Korbinian Strimmer for helpful comments on a first draft. All errors are my own. All analyses, maps and figures were done in *R* (R Development Core Team 2008), with the additional packages *vcd* (Meyer et al. 2006) and *Design* (Harrell 2001). I am especially grateful to Hans-Jörg Bibiko for greatly improving R's map tools.

References

Agresti, Alan (2002). *Categorical data analysis.* New York: Wiley-Interscience.

Baayen, R. Harald (in press). *Analyzing linguistic data: a practical introduction to statistics.* Cambridge: Cambridge University Press.

Beaton, A. C. (1968). *A Grammar of the Fur Language.* Khartoum: Sudan Research Unit, University of Khartoum.

Bickel, Balthasar (in press). A refined sampling procedure for genealogical control. *Sprachtypologie und Universalienforschung* #, ## - ##.

Bickel, Balthasar, Kristine Hildebrandt & René Schiering (in press). The distribution of phonological word domains: a probabilistic typology. In Janet Grijzenhout & Bariş Kabak (eds.) *Phonological Domains: Universals and Deviations.* Berlin: Mouton de Gruyter.

Bickel, Balthasar & Johanna Nichols (1996ff). The AUTOTYP database. Electronic database; http://www.uni-leipzig.de/~autotyp.

Bickel, Balthasar & Johanna Nichols (2003). Typological enclaves. Paper presented at the 5th Biannual Conference of the Association for Linguistic Typology, Cagliari, September 18; available at http://www.uni-leipzig.de/~autotyp/download.

Bickel, Balthasar & Johanna Nichols (2006). Oceania, the Pacific Rim, and the theory of linguistic areas. *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society*, [PDF available at http://www.uni-leipzig.de/~bickel/research/papers].

Bickel, Balthasar & Johanna Nichols (in press). The geography of case. In Andrej Malchukov & Andrew Spencer (eds.) *The Handbook of Case*. Oxford: Oxford University Press.

Blevins, Juliette (2004). *Evolutionary phonology : the emergence of sound patterns.* New York: Cambridge University Press.

Bybee, Joan (1988). The diachronic dimension in explanation. In John A. Hawkins (ed.) *Explaining language universals*, 350 – 379. Oxford: Blackwell.

Chaubey, Gyaneshwer, Mait Metspalu, Toomas Kivisild & Richard Villems (2006). Peopling of South Asia: investigating the caste-tribe continuum in India. *BioEssays* 29, 91 - 100.

Comrie, Bernard (2005). Alignment of case marking. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The world atlas of language structures*, 398 – 405. Oxford: Oxford University Press.

Corbett, Greville G. (2005). Number of genders. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The world atlas of language structures*, 126 – 129. Oxford: Oxford University Press.

Croft, William (2000). *Explaining language change: an evolutionary approach.* Harlow: Longman.

Cysouw, Michael (2003). Against implicational universals. *Linguistic Typology* 7, 89-110.

Cysouw, Michael (2005). Quantitative methods in typology. In Gabriel Altmann, Reinhard Köhler & R. Piotrowski (eds.) *Quantitative linguistics: an international handbook*, 554 – 578. Berlin: Mouton de Gruyter.
Dryer, Matthew S. (1989). Large linguistic areas and language sampling. *Studies in Language* 13, 257 – 292.
Dryer, Matthew S. (2000). Counting genera vs. counting languages. *Linguistic Typology* 4, 334 – 350.
Dryer, Matthew S. (2002). Case distinctions, rich verb agreement, and word order type (comments on Hawkins' paper). *Theoretical Linguistics* 28, 151 – 157.
Dryer, Matthew S. (2005a). Order of relative clause and noun. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The world atlas of language structures*, 366-369. Oxford: Oxford University Press.
Dryer, Matthew S. (2005b). Order of subject, object, and verb. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The world atlas of language structures*, 330 – 334. Oxford: Oxford University Press.
Everitt, Brian S. & Torsten Hothorn (2006). *A handbook of statistics using R.* Boca Raton, FL: Chapman & Hall.
Forster, J.J., J.W. McDonald & P.W.F. Smith (2003). Markov Chain Monte Carlo exact inference for binomial and multinomial logistic regression models. *Statistics and Computing* 13, 169 – 177.
Good, Phillip I. (2006). *Resampling methods : a practical guide to data analysis.* Boston: Birkhèauser.
Greenberg, Joseph H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.) *Universals of Language*, 73 – 113. Cambridge, Mass.: MIT Press.
Greenberg, Joseph H. (1978). Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.) *Universals of human language I: method and theory*, 61 – 92. Stanford: Stanford University Press.
Greenberg, Joseph H. (1995). The diachronic typological approach to language. In Masayoshi Shibatani & Theodora Bynon (eds.) *Approaches to language typology*, 143 – 166. Oxford: Clarendon.
Hall, Christopher J. (1988). Integrating diachronic and processing principles in explaining the suffixing preference. In John A. Hawkins (ed.) *Explaining language universals*, 321 – 349. Oxford: Blackwell.
Harrell, Frank E. (2001). *Regression modeling strategies.* New York: Springer.
Haspelmath, Martin (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18, 180 – 205.
Hawkins, John A. (2004). *Efficiency and complexity in grammars.* Oxford: Oxford University Press.
Jakobson, Roman (1931). K xarakteristike ėvrazijskogo jazykovogo sojuza. *Selected Writings 1*, 144 – 201. The Hague: Mouton 1970.
Janssen, Dirk, Balthasar Bickel & Fernando Zúñiga (2006). Randomization tests in language typology. *Linguistic Typology* 10, 419 – 440.
Johnson, Keith (in press). *Quantitative methods in linguistics.* London: Blackwell.
Justeson, J.S. & L.D. Stephens (1990). Explanation for word order universals: a log-linear analysis. *Proceedings of the XIV International Congress of Linguists*. Berlin: Mouton de Gruyter.
Kirby, Simon (1999). *Function, selection, and inateness.* Oxford: Oxford University Press.
Maddieson, Ian (2006). Correlating phonological complexity: data and validation. *Linguistic Typology* 10, 106 – 123.
Manly, Bryan F. J. (2007). *Randomization, bootstrap and Monte Carlo methods in biology.* Boca Raton, FL: Chapman & Hall/ CRC.
Masica, Colin (2001). The definition and significance of linguistic areas: methods, pitfalls, and possibilities (with special reference to the validity of South Asia as a linguistic area). In Peri Bhaskararao & Karumuri Venkata Subbarao (eds.) *Tokyo Symposium on South Asian languages: contact, convergence, and typology [= The Yearbook of South Asian Languages and Linguistics 2001]*, 205 - 267. New Delhi: Sage Publications.
Maslova, Elena (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4, 307 – 333.

Maslova, Elena & Tatiana Nikitina (2007). Stochastic universals and dynamics of cross-linguistic distributions: the case of alignment types Ms. Stanford University, http://www.stanford.edu/~emaslova/Publications/Ergativity.pdf [accessed Dec 1, 2007].
Meyer, D., A. Zeileis & K. Hornik (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software* 17, 1 - 48.
Nasidze, I., T. Sarkisian, A. Kerimov & M. Stoneking (2003). Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. *Human Genetics* 112, 255 – 261.
Nichols, Johanna (1992). *Language diversity in space and time.* Chicago: The University of Chicago Press.
Nichols, Johanna (1993). Ergativity and linguistic geography. *Australian Journal of Linguistics* 13, 39 – 89.
Nichols, Johanna (1997). Modeling ancient population structures and population movement in linguistics and archeology. *Annual Review of Anthropology* 26, 359 – 384.
Nichols, Johanna (1998). The Eurasian spread zone and the Indo-European dispersal. In Roger Blench & Matthew Spriggs (eds.) *Archeology and language II: archeological data and linguistic hypotheses*, 220 - 266. London: Routledge.
Nichols, Johanna (2003). Diversity and stability in language. In Richard D. Janda & Brian D. Joseph (eds.) *Handbook of Historical Linguistics*, 283 – 310. London: Blackwell.
Perkins, Revere D. (2001). Sampling procedures and statistical models. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.) *Language typology and language universals, vol. 1*, 419 – 434. Berlin: Mouton de Gruyter.
R Development Core Team (2008). *R: a language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing (www.r-project.org).
Rijkhoff, Jan (2002). *The noun phrase.* Oxford ; New York: Oxford University Press.
Rootsi, S., L.A. Zhivotovsky, M. Baldovic, M. Kayser, I.A. Kutuev, R. Khusainova, M.A. Bermisheva, M. Gubina, S. Fedorova, A.M. Ilumäe, E.K. Khusnutdinova, L.P. Osipova, M. Stoneking, V. Ferak, J. Parik, T. Kivisild, P.A. Underhill & R. Villems (2007). A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *European Journal of Human Genetics* 15, 204 - 211.
Siewierska, Anna (1996). Word order type and alignment. *Sprachtypologie und Universalienforschung* 49, 149 – 176.
Thomason, Sarah Grey & Terrence Kaufman (1988). *Language contact, creolization, and genetic linguistics.* Berkeley: University of California Press.
Widmann, Thomas & Peter Bakker (2006). Does sampling matter? a test in replicability, concerning numerals. *Linguistic Typology* 10, 83 - 95.
Zamar, David, Brad McNeney & Jinko Graham (2007). elrm: software implementing exact-like inference for logistic regression models. *Jounal of Statistical Software* 21, 1 - 18.

MAP 1

MAP 2

MAP 3

FIGURE 1

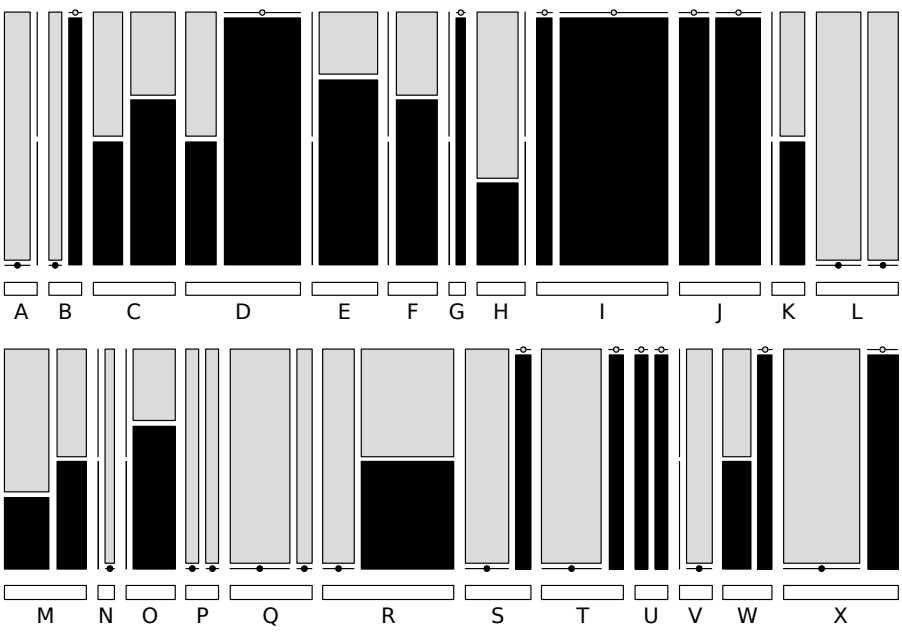

FIGURE 2

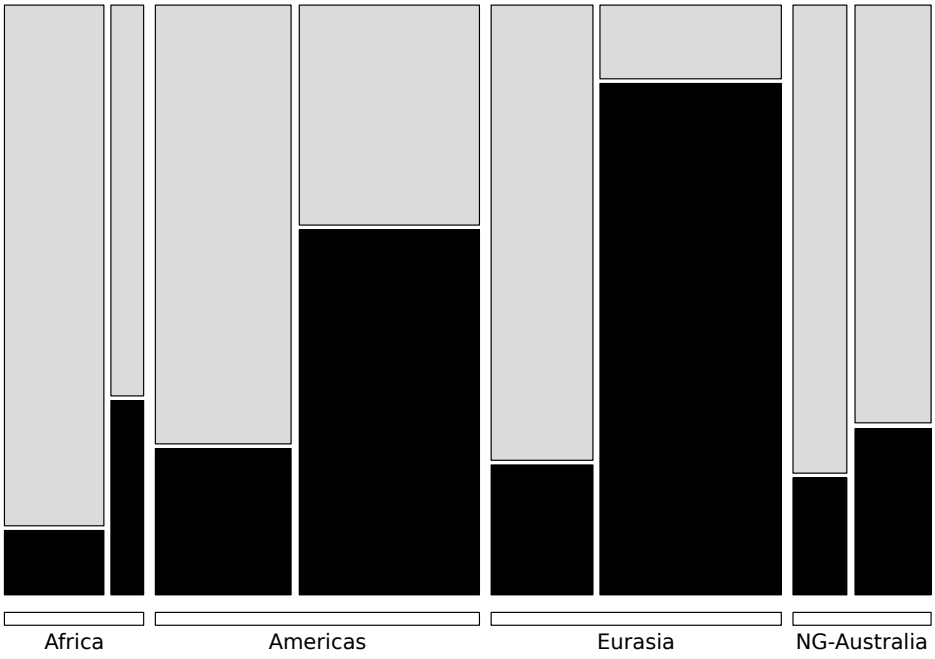

FIGURE 3