Distributional biases in language families*

Balthasar Bickel
University of Leipzig

"Stability or instability [...] is a matter of competing forces."
(Nichols 2003:283)

1 Introduction

In her programmatic paper on "diversity and stability in language" (2003), Johanna Nichols sketches a theory of diachronic stability. One of the key insights of this theory is that degrees of stability are not self-contained indices of language change but the result of competing forces, such as diachronic replication, borrowing, substratal effects, and universals. In this chapter I develop and discuss methods for estimating the role of these forces on the basis of statistical analyses of synchronic typological datasets.

At first sight, the stability of a typological variable seems established when, synchronically, all or nearly all members of a family have the same value on that variable (e.g. all have postpositions), and this is so for many families. But such a bias can have very different sources. The bias can be caused by genealogical stability: daughter languages have or tend to have the same values because they inherited these from the proto-language. Alternatively, the exact opposite of this is also possible, and the bias can be caused by genealogical instability: daughter languages have the same values because they all changed in the same direction. Such a

* Parts of this paper were presented at the 43rd Annual Meeting of the *Societas Linguistica Europaea*, September 3, 2010 in Vilnius, and I also discussed some of the methodological issues presented here in my course on quantitative methods in typology at the *DGfS-CNRS Summer School on Linguistic Typology*, Leipzig, August, 15 - 17, 2010. I am grateful to both audiences for questions and comments. Special thanks go to Taras Zakharko for many very useful comments, including the suggestion to use the Laplace estimator in Section 5. Some of the ideas go back to Bickel (2008a), where I used the term 'skewing' instead of 'biases'. The term 'skewing' is unsuitable because of possible ambiguities with 'skewing' in the sense of lopsided distributions. A previous version of the current paper was circulated in January 2010. Apart from matters of exposition and exemplification, the most important change is that families with few members are now treated like isolates and not like large families (in response to a question raised by Alena Witzlack-Makarevich). All computations presented were carried out in R (R Development Core Team 2011). The research presented here was supported by Grant Nr. II/83 393 from the Volkswagen Foundation.

change could be caused for example by repeated diffusion effects (keeping quirks like relative pronouns at bay worldwide) or by developing cognitively favored structures (e.g. by favoring agent-before-patient constituent orders). If a bias can result from both stability and instability, the question arises how one can tell these possibilities apart.

In the following, I propose an answer to this question in terms of what I call the *Family Bias Theory*. I first introduce, illustrate and motivate the basic ideas of the theory based on univariate distributions (Sections 2 – 4). In Section 5 I explore possible extrapolations to small families and isolates, and in Section 6 I discuss extensions to multivariate distributions. Section 7 compares the Family Bias Theory with classical approaches of genealogically balanced sampling and Section 8 concludes the chapter by discussing implications for future research.

2 The Family Bias Theory: the basic ideas

If one surveys language families (in the sense of a genealogical unit established by the Comparative Method), one quickly notices that a family can be uniform or near-uniform (e.g. all or most members of the family have a dual), or it can be diverse (some members have, others don't have a dual, as in Indo-European). In other words, each family may or may not show a bias towards a given feature. The extent and significance of this bias can be established by standard statistical techniques, such as χ^2-tests.

Biases of this kind may show up at any taxonomic level and with any time depth: for example, whereas there is a bias towards verb-final clause structures at the stock level in many old stocks of New Guinea, the same bias is found only at shallower levels in branches of Indo-European, where the stock as a whole is much more diverse. In the following, I use the term 'family' as a generic term for any taxonomic level, and reserve 'stock' for the highest proven taxon (following Nichols 1992, 1997a).

Families can be biased with regard to any kind of typological variable: they can be biased by favoring the presence or the absence of a feature or a feature set, a certain value or interval on a continuous variable, or some complex constellation of such characteristic. In the following, I use the symbol *F* to subsume all these possibilities.

The basic proposal of the Family Bias Theory is that the synchronic distribution of a typological variable across families reflects distinct historical scenarios. Specifically, the nature of biases across families reflects two different scenarios:

A. **Directional Family Bias**: If there are significantly more families that are biased towards *F* than towards *non-F*, this reflects **universal pressure** in the sense that the development and maintenance of *F* is universally preferred over the development and maintenance of *non-F*. The total proportion of families with a bias (in either direction) as against families that are diverse indicates the lower bounds of how strong the universal pressure is.

B. **Non-Directional Family Bias**: If there are significantly more families that are biased rather than diverse, but the bias is undirected, i.e. an equal proportion of families are biased towards *F* or *non-F*, the variable tends to be **genealogically stable** in the sense that *F* tends to be unconditionally and faithfully replicated and that changes from *F* to *non-F* or from *non-F* to *F* tend to be disfavored.

In this, 'universal pressure' refers to any principle suspected to shape the structure of languages: preferred structures tend to universally develop more easily than dispreferred structures, and they tend to be universally maintained more persistently than dispreferred structures.[1] There are many ways of how universal pressure can work: for example, by favoring the most frequent patterns in discourse or those that are easiest to process in comprehension, or they can be based on more abstract principles like iconicity or paradigm symmetry. Also, universals may operate as selectors of variants in language change or as pathways of change themselves. In the following I gloss over all these differences and do not discuss the working mechanisms and ultimate causes of universal pressure. My interest is only in determining the kinds of effects on the synchronic distribution of families biases that one can expect if some kind of universal pressure is at work.

The presence of a significant direction in family biases (Scenario **A**) is independent of the relative proportion of diverse vs. biased families. Diverse families do not provide evidence for or against a universal. On the one hand, a family may be diverse because the proto-language complied with the universal (i.e. had the preferred pattern), and some daughter languages moved away from that pattern. In this case, the family would be counter-evidence against a universal. On the other hand, a family may be diverse because of developments in exact opposition to this and in line with the universal: the proto-language may not have complied with the universal, and some daughter languages have changed towards the preferred pattern. This would favor the hypothesized universal. Unless we know the relevant patterns in the proto-language for sure (which we usually don't), both possibilities are equally likely.

What the proportion of diverse families does indicate, however, is the strength of universals. If despite a significant direction of the bias (many more biases towards *F* than towards non-*F*), there is a relatively large proportion of diverse families, this suggests that *F* tends to change relatively quickly, and that the universal pressure consists in a relatively low probability of change towards *F*. This probability must still be higher than the probability of change in the opposite direction (for else there would not be evidence for a universal direction in biases across many families). For example, there could be weak effects in language processing that favor certain patterns, but the probability that the effects leave a trace in language change are relatively small, and so it takes many generations in many families for the effects to become visible in extant distributions.

If the proportion of diverse families is small and there is a significant direction in the bias, this means that universal pressure is very strong: if a language deviates from the preferred pattern, there is strong pressure to 'correct' this, and this quickly leads to uniform or near uniform daughter languages, all with the preferred pattern. Conversely, once the preferred pattern is established, there is strong universal pressure not to loose the pattern.

Because, as noted, diversity in families can arise both with and without universal pressure, the proportion of diverse families only approximates the lower bounds of the strength of the universal. The real strength is underestimated to the extent that diverse families in fact contain

[1] This is in fact nothing but a restatement of the by-now classic view of universals as diachronic laws of type preference; cf., among others, Bybee (1988), Hall (1988), Nichols (1992, 2003), Greenberg (1995), Haspelmath (1999), Maslova (2000a), Blevins (2004), Bickel (2007), Maslova & Nikitina (2007).

incipient effects of a universal. However, as far I can see, this extent cannot be estimated on the basis of synchronic surveys.

If the proportion of diverse families is small but there is no significant direction, this is Scenario **B**: a trend towards high copy fidelity from generation to generation. Finally, a typological variable can of course also show no significant trend in family biases, i.e. neither a preference for families to be biased vs. to be diverse (Scenario **B**), nor a preference towards *F* vs. *non-F* within biased families (Scenarios **A**). Such a situation does not suggest any particular pattern, and the current distribution is mostly the result of chance events in language change.

In the following, I first illustrate the two scenarios in Section 3 and then provide evidence and argumentation for the theory in Section 4. For illustration, I rely on the genealogical taxonomy of Nichols & Bickel (2009), and I only consider families with several representatives in typological databases. Extrapolations to isolates or under-sampled stocks are discussed in Section 5.

3 Illustrations of the theory

In the illustrations I concentrate on families with at least 5 members and as a criterion for what I count as a significant bias in a family, I choose a rejection level of 10% in a χ^2 permutation test. These thresholds yield *p*-values that match the intuition that complete consistency in a family (i.e. 5 out of 5 members have *F*) represents a linguistically interesting bias (with $p \approx .06$). However, not much depends on these parameters, and the results are similar if one chooses a lower rejection level or limits surveys to families with more members (or both).

The examples in the following are only meant to show how the two scenarios play out in terms of data distributions. Specifically, if a dataset supports Scenario A, this suggests the existence of universal pressure and invites further research. But it cannot and does not demonstrate or prove such pressure. This can only be done by reversing the procedure, i.e. by first developing a well-motivated and fully-fledged causal model of how a suspected universal could have a systematic impact on language change, and then test the resulting hypothesis against large datasets (in fact, larger than what I have here access to for illustrative purposes). In other words, the illustrations can at best be indicative of statistical trends, and like all statistical trends, they may or may not reflect real causal chains.

3.1 A Scenario A example: A-before-P order

A good example of a directional family bias (Scenario **A**) is Greenberg's Universal 1 (Greenberg 1963), the worldwide trend towards placing agents before patients in simple clauses. To assess the distribution of this property I defined a binary variable capturing whether a language has a rigid or at least dominant A-before-P order in contrast to a language that allows various orders or even favors P-before-A orders. I then applied this to a dataset merging Dryer's (2005b) WALS data and data from the AUTOTYP database, covering 1,372 languages in total.[2] Each stock was

[2] For the AUTOTYP data, see `http://www.uni-leipzig.de/~autotyp`. Merging the data is justified by the fact that there are only 3% mismatches in the 558 languages for which there is information in both databases. When there were mismatches, I chose the AUTOTYP coding.

tested for whether it is biased towards an A-before-P or towards the opposite (i.e. P-before-A or flexible order). A stock counts as biased if there is a significant preference for one of the two options.

Determining such biases in stocks suggests that of the 59 large stocks in the dataset, 39 (66%) are biased towards an A-before-P order and only 2 (3%) are biased towards the opposite. The two opposite biases (Algic and Iroquoian) are both towards flexible orders, not towards a rigid P-before-A order. The remaining stocks (18, corresponding to 31%) are diverse and show no significant bias in any direction (e.g. Cariban has a mix of languages with flexible, A-before-P and P-before-A orders).

These frequencies suggest that there is a significant and large preference for families to be biased towards an A-before-P order (exact binomial test, one-tailed $p < .001$, $\hat{\pi} = .95$).[3] Under the assumption of the Family Bias Theory, this would suggest that there is universal pressure for families to keep A-before-P orders if the proto-language already had this, or to develop such orders if the proto-language did not have such an order. Since we found the biases at the level of stocks, this means that universal pressure must have been strong enough to affect language change within the time-depth of stocks (by keeping languages from changing away from A-before-P orders and by favoring changes towards A-before-P orders). The minimum strength s of the effect can be estimated from the proportion of biased among all families, which is at least $\hat{s} = .69$.

3.2 A Scenario B example: coding of property concepts in predicate position

Non-directional family bias (Scenario **B**) can be illustrated by the distribution of how languages code predicative adjectives. Stassen (2005) defines three types, verbal (i.e. verb-like), nonverbal and mixed coding of property concepts in predicative function. 'Mixed' means that languages use both strategies, either differentiated by function or by lexical classes (e.g. verbal coding is used to predicate temporary properties and nonverbal coding is used to predicate permanent or intrinsic properties).

Stassen's (2005) database contains 18 sufficiently large families (i.e. families with at least 5 members each). Of these, 13 (72%) are biased in some direction, and this proportion exceeds what one would expect by chance alone (exact binomial test, one-tailed $p = .048$). Within biased families, 6 families prefer nonverbal, 4 verbal and 3 mixed coding. These proportions are statistically fairly close to a uniform (i.e. ⅓ each) distribution, and so there is no evidence for any one type being universally preferred ($\chi^2 = 1.08$, $p = .69$). Under the assumption of the Family Bias Theory such a finding suggests that the way predicative property concepts are coded is genealogically stable at the time depth of the assumed genealogy. Almost three quarters of the stocks in the database tend to have a consistent type throughout the family, suggesting a strong bias towards diachronic inertia: if the proto-language had a specific type, this tends to survive

[3] I use binomial tests here although Poisson (log-linear) modeling might eventually be more appropriate since it is plausible to think of family biases as Poisson processes; see Cysouw (2010b) for some arguments for Poisson processes as underlying typological distributions. None of the results reported here depends on the decision, as p-values are in the same ballpark anyway. Note that I use the symbol π for the probability of an event, in order to avoid confusion with p-values (the probability of a test statistic under the null hypothesis).

all splits and branchings. Only about one quarter of the stocks in the database are diverse so that within them, some branches must have lost or innovated a type.

Of course, there might be additional, e.g. areal or structural factors that favor one or the other types. From Stassen's (2005) map it looks like South-East Asia for example is an area with a very strong preference for the verbal coding type, but it is not clear where the boundaries of this area would be in this case: in one sense, it extends all over the Pacific Rim (Nichols 1992). But this is contradicted by many languages with nonverbal types in South America, Australia and the Papuan region. Anyway, as far as I can see, there is so far no statistical signal in any clear direction.

An additional and in fact more severe problem is that the total number of families that have enough members for estimating biases is relatively low: there are only 18 stocks with more than 5 members each, but critical statistical signals could come from smaller families and isolates. I will return to this problem and suggest a solution in Section 5. Before this, however, I wish to further discuss and substantiate the core claims of the Family Bias theory.

4 Evidence for the theory

The central claim of the Family Bias Theory is that a directional bias across families reflects some driving factor (universal pressure, as per Scenario A). The alternative to this view would be to hypothesize that a directional bias reflects not some driving factor but instead faithful inheritance, i.e. extremely stable distributions (as per Scenario B). As a result, not only non-directional but also directional family biases would ultimately be caused by diachronic inertia, a general reluctance to change over time. In the example of the A-before-P order, this would mean that most families consistently have A-before-P orders not because this order is universally privileged but because speakers faithfully copy this order from their parental languages and most parental languages just happened to have had A-before-P order.

Technically, the difference between these two hypotheses boils down to differences in probabilities of change, as in spelled out in (1), where the succession symbol represents diachronic change and F again represents some typological characteristic:

(1) *Two possible hypotheses explaining directional family bias:*
   a. $\pi(\text{non-}F \succ F) > \pi(F \succ \text{non-}F)$
   b. $\pi(\text{non-}F \succ F) \approx \pi(F \succ \text{non-}F) \approx 0$

One of the points of Maslova (2000a) is that it is nearly impossible to decide between these two hypotheses. By contrast, the key claim of the Family Bias Theory is that (1b) needs to be rejected as unrealistic. A similar point was made by Johanna Nichols in her 2002 plenary address to the Linguistic Society of America (Nichols 2002), and in the following I substantiate the arguments and evidence for this.

Let us assume, for the sake of the argument, that hypothesis (1b) is correct, and that accordingly, a directional family bias in a distribution D results by and large from faithful replication within each family. If this is so, then the distribution in the current generation $D(G_0)$ must resemble the distribution in the previous generation, $D(G_{k+1})$. Unless there was some driving factor before G_{k+1}, all $D(G_k)$ must reflect $D(G_{k+1})$ until k spans the entire history of the human

language faculty. Then, D can be said to be super-stable over very deep time, and from this, we can predict that changes in D are all the more unlikely within short time intervals. Now, all reconstructible time intervals are relatively short — up to about 6-8,000 years, the age of demonstrable families. Therefore, if a variable is super-stable, we except to be able to observe almost no changes in the known history of $D(G_0)$. As a result, most observable families should be uniform since each case of a non-uniform family necessarily represents at least one case of change. Given this, the empirical question is whether one can observe more cases of change in $D(G_0)$ than what would be expected if $D(G_0)$ is the sole result of faithful replication, defined as a small probability of change π in (1). The more we observe cases of change beyond what small values of π allow, the less such values become, and this would disfavor Hypothesis (1b).

To find out, I first computed the minimum number of changes attested in each known family for a large set variables. This corresponds to the number of unique values ('types', 'levels') in each family, minus 1: if a stock has two different values in one variable, there must have been at least one case of change (regardless of how the tokens are actually distributed, e.g. as 1:9 or a 5:5 ratio), e.g. from 'A' to 'B' or vice-versa. If a family has three different values, there must have been at least two cases of change, e.g. from 'A' to 'B' and to 'C', and so on. There could of course always have been more cases of change (in parallel or in sequence), but the logical minimum of observable changes equals the number of unique values minus one:

(2) $\min(C_F) = k_F - 1,$

where C_F represents changes in variable F and k_F the number of levels (types) of F.[4]

I computed this minimum for a total of 386 variables taken from the AUTOTYP and WALS databases, requiring that the variable is coded for at least 10 families that each are represented by at least two members (since isolates or families represented by a single member do not allow counting cases of change). The variables are of various kinds covering almost all parts of grammar and phonology, and they include many alternative ways of coding, e.g. both a binary and a 6-way breakdown of basic word order, and various other versions of this. For current purposes, I treat scalar variables (e.g. on the size of the vowel inventory) in the same way as categorical ones. This is justified by the fact that from the point of view of diachrony, a change from one point on a scale to another, e.g. from 5 to 6 vowels, is as discrete a change as, say, the development of a tone opposition from laryngeal setting contrasts.

I then tested for each variable whether the observed minimum number of changes, i.e. $\min(C_F)$, exceeds what can be expected under the assumption of a given probability of change π, letting π represent various assumptions ranging from $\pi = 0$ to $\pi = 1$ (at increments of .01), and assuming, for the sake of the argument, that the current distributions are the sole result of faithful represent (as per Hypothesis 1b). As a criterion for what qualifies as an unexpected excess of $\min(C_F)$ under a given value of π, I use a .05 rejection level of the null hypothesis that the observed proportion of $\min(C_F)$ does not exceed π in a one-sided binomial test. If the observed proportion significantly exceeds what is expected under a given value of π, this means that the actual probability of change must be higher than π.

[4] And note that counting the minimum in this way favors the hypothesis in (1b) because it systematically underestimates the probability of change π.

With binary variables, the observed proportion of $\min(C_F)$ can be directly computed by dividing $\min(C_F)$ by the number of families in the database, since each family corresponds to at least one 'opportunity' for the variable to change, e.g. from type 'A' to type 'B' or vice-versa (always limiting our attention, as before, to the *minimum* number of changes that is logically possible). For example, with π = .15, it is unexpected (under a binomial test) to find a minimum of 20 cases of change in 50 families if the variable is binary. But if the variable defines three instead of two values, each family allows for at least two possible changes: from 'A' to 'B' or to 'C', and then it is no longer unexpected to find at least 20 cases of change. In general, for a variable that defines k types, the (minimum) number of opportunities for change is

(3) $$\min(O_F) = (k_F - 1) \cdot N(families)$$

Therefore, I tested whether the proportion $\frac{\min(C_F)}{\min(O_F)}$ is expected under a given probability π.

The result of these tests for the 386 variables is summarized in Figure 1. The assumed values of π only reach a more substantial match with observed numbers of changes if $\pi > .10$, starting with a proportion of 25%, and they only reach full coverage if $\pi \geq .58$. This is far above what Hypothesis (1b) allows for.

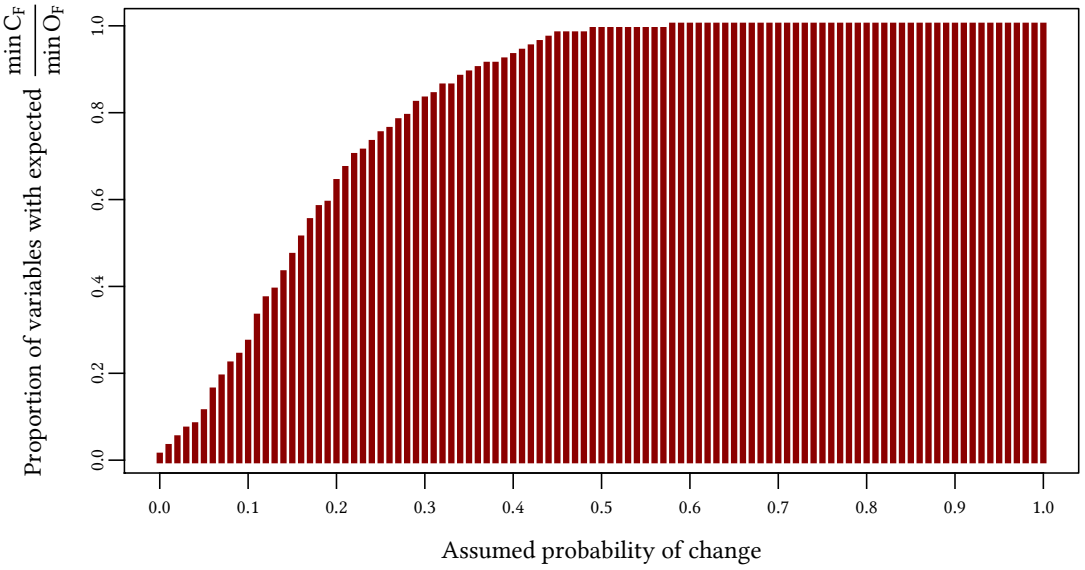


Figure 1: Proportion of variables for which the observed minimum numbers of change is statistically expected under assumed probabilities of change π.

However, (1b) is difficult to maintain even for those variables with low numbers of observed changes and that are therefore compatible with the assumption of low values of π. This becomes evident in Figure 2, which plots the mean entropies of those variables for which the observed numbers of changes is statistically expected under a given value of π. Entropies (designated H) are a standard estimate of the extent to which the distribution of values in a variable is

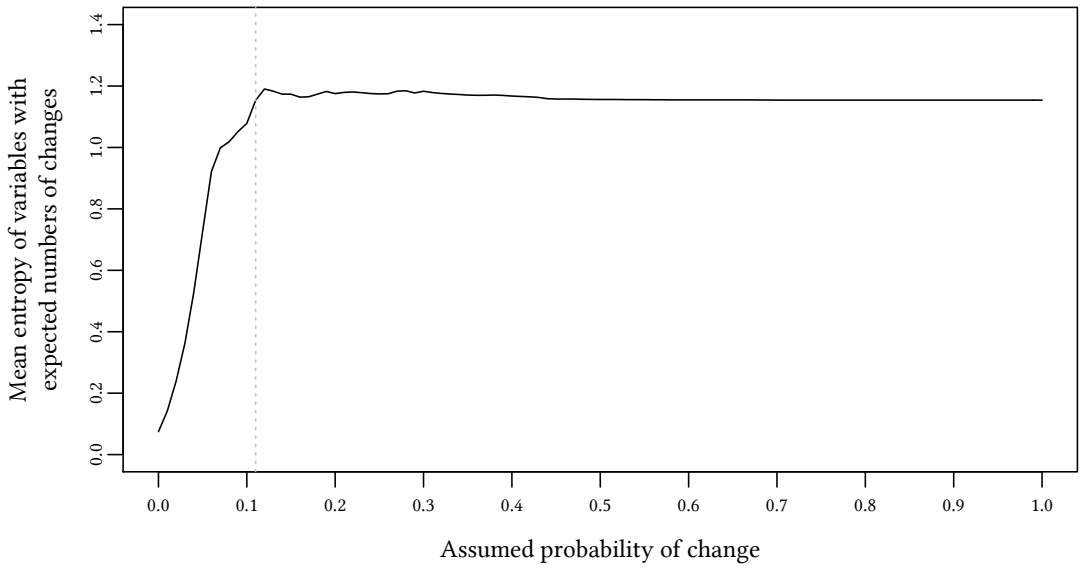


Figure 2: Mean entropies of those variables for which the observed minimum numbers of change is statistically expected under a given probability of change π.

biased (low entropy) rather than uniform (high entropy).[5] As indicated by the grey dotted line, the mean entropies reach their overall characteristic values (with mean $\hat{H}$ = 1.13) only with $\pi > .10$, i.e. only with variables for which the observed number of changes is expected under assumed probabilities higher than .10. With $\pi \leq .10$, variables tend to have considerably lower entropies, reflecting strong biases towards one value. In other words, in this range of π there are, on average, only very few variables with more balanced distributions (such as the coding of property concepts in predicative function, as reviewed in Section 3).

Strong biases (low entropies) are characteristic of *rara* vs. *universalia* oppositions. Table 1 illustrates this for those variables for which the expected number of changes is covered at π = .01, i.e. at a value of π that would be closest to what is hypothesized in (1b). The presence of 'have'-perfects and of tonal case are well-known areal *rara* in Europe and Africa, respectively, and both have relatively shallow histories — i.e. the exact opposite of what (1b) would predict. All other variables reflect very strong universal pressure in favor of some feature (independent subject pronouns, interrogative/declarative distinctions) or against some feature (stem flexivity conditioned by negation markers, or various co-exponence types of such markers[6]). This is fully in line with the hypothesis in (1a): for example, it seems much more common to develop and maintain an interrogative/declarative distinction than to loose it. But it is difficult to explain if

[5] Formally, the (Shannon) entropy H of a variable V with levels $v_i \in \{v_{1\ldots k}\}$ and associated probabilities π_{v_i} is $H(\mathrm{V}) = -\sum_{i=1}^{k} \pi_{v_i} \log_2(\pi_{v_i})$. $H(\mathrm{V})$ is zero if there is a maximum bias towards a single level, e.g. with $\pi_{v_1} = 1$, $\pi_{v_2} = 0$, and $\pi_{v_3} = 0$; $H(\mathrm{V})$ reaches its maximum in uniform distributions, e.g. with $\pi_{v_1} = \frac{1}{3}$, $\pi_{v_2} = \frac{1}{3}$, and $\pi_{v_3} = \frac{1}{3}$. I estimate π using the Maximum Likelihood method, i.e. from the empirical frequencies.

[6] This has been noted by Bickel & Nichols (2005) for the co-exponence of other categories as well.

the development in either direction has a very low probability (as 1b would predict). For other variables in the range $.01 \leq \pi \leq .10$, the picture is similar to what is illustrated by Table 1: they tend to be heavily biased and reflect a *rara* vs. *universalia* distribution. Such biases are likely to result from strong areal diffusion or universal pressure — so strong in fact that the relevant choice is likely to establish itself very quickly, and that once the choice is made, languages refrain from undoing it and families look almost completely homogenous. This reflects the scenario hypothesized in (1a) and is not consistent with (1b). Thus, rather than suggesting faithful replication, extremely low numbers of known changes seem to point to very strong effects of some driving factor (*pace* Parkvall 2008, Wichmann & Holman 2009, or Bakker et al. 2009).

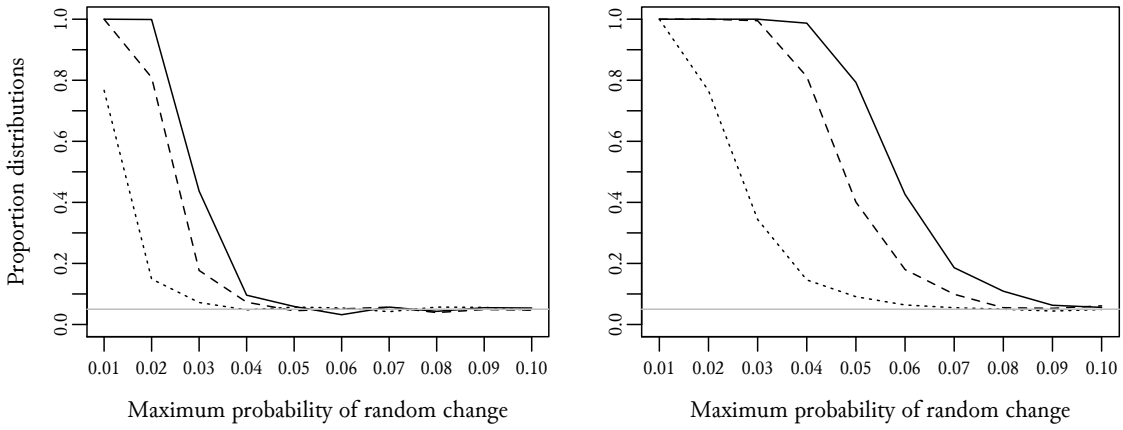
Variable (and data source)	Changes $\min(C_F)$	Opportunities $\min(O_F)$	Entropy $\hat{H}$	Ratio of values
Interrog./decl. distinction (Dryer 2005c)	1	89	0.01	841:1
Indep. subject pronouns (Daniel 2005)	0	31	0.07	258:2
Tonal case (AUTOTYP and Dryer 2005d)	3	91	0.07	698:6
Stem flexivity condit. by NEG (AUTOTYP)	0	40	0.12	141:1:1
'Have'-perfect (Dahl & Velupillai 2005)	1	15	0.35	101:7
Co-exponent type of NEG (AUTOTYP)	4	234	0.60	185:5:3:1:1:1:1:1:1:1

Table 1: Variables for which the minimum number of changes does not exceed what is expected under $\pi = .01$ (in increasing order of entropy)

There is one further piece of evidence against Hypothesis (1b): already with $\pi = .05$ and certainly with $\pi = .10$, it is virtually impossible for typological distributions to persist over deep time in such a way that what one observes now is similar to what was there many generations ago. This can be shown by computer simulations. I set up datasets with 1,300 fake languages (approximating the size of the largest available real databases) with fake codings for a binary typological variable. The codings represent distributions ranging from 1%:99% to 20%:80% to 40%:60%. Each such distribution was then sent through a number of generations. In each generation there was a certain probability threshold π (ranging from $\pi = .01$ to $\pi = .10$, at increments of .01) below which a random subset of languages would change from one state to the other, with no preferred direction of change. Choosing random subsets below π rather than at π is motivated by the assumption that language change is constrained by maximum probabilities but does not operate at a constant rate; in addition, the method favors Hypothesis (1b) since change does not always operate 'at full speed' as it were. After the distributions went through all generations, I tested whether the initial distribution was still detectable using a two-sided binomial test. This procedure was repeated 1,000 times, allowing to compute the proportion of simulations in which the original distribution was still detectable, and from this an estimate of the overall probability of successful detection.

Figure 3a reports the results for 100 generations, and Figure 3b for 50 generations. If we assume an average lifespan of languages of about 1,000 years, 100 generations reflect a low estimate of the age of human language, i.e. a time when major innovations that are likely to depend on language use, such as ornamentation, pigment processing, and long-distance trading,

become well attested in the archeological record (McBrearty & Brooks 2000). 50 generations reflects an unrealistically low estimate, viz. a time when modern symbolic behavior has spread even well outside Africa.



(a) After 100 generations of change

(b) After 50 generations of change

Figure 3: Proportion of 1,000 simulated distributions that are still detected after 100 (a) or 50 (b) generations of random change at given maximum probabilities (solid line: initial distribution 1%:99%, dashed line 20%:80%, dotted line 40%:60%; grey line: .05 probability threshold of detecting a an initial distribution)

The findings suggest that already at π = .05, the probability of detecting even the most heavily biased initial distribution (1%:99%, plotted as a solid line) starts to no longer exceed .05 after 100 generations (as indicated by the grey horizontal line in the figure); even under the shorter scenario of 50 generations, the most heavily biased distribution reaches the .05 probability threshold just before π = .10, the level that was noted above as the minimum at which an appreciable proportion of variables (about 25%) begins to show numbers of change that are statistically expected. Thus, even under an unrealistically short life span of human language, the minimum values of π that begin to be consistent with the number of known changes is far too high for allowing the long-time persistence of typological distributions required by (1b). In other words, realistic values of π are so high that no synchronic distribution can be accounted for by faithful replication over deep time.

To summarize, there are three pieces of evidence against the hypothesis of extremely low probabilities of change in typological variables (i.e. Hypothesis 1b): first, we tend to find many more cases of change than what extremely low probabilities of change would lead us to expect. Second, those few variables for which extremely low probabilities of change would in principle match the observed number of changes, tend to display an extreme *rara* vs. *universalia* distribution, and this fits better with very unequal probabilities of change (1a) than with equal probabilities (1b). Third, even if probabilities of change were as low as .10, they would still be too high for typological distributions to remain stable over the entire lifespan of the human

language faculty.

Taken together, these three pieces of evidence make the hypothesis in (1b) an unlikely explanation of directional family biases. This supports the alternative in (1a) and thereby the core claim of the Family Bias Theory: that directional family biases reflect the pressure of some driving factor and that high degrees of genealogical stability can only explain non-directional family bias but not also directional family bias.

5 The problem of small families

The Family Bias Theory provides a systematic diachronic interpretation of synchronic typological distributions. However, like all diachronic interpretations it has a natural limit when confronted with small families: it is difficult to estimate a bias or reconstruct forms if one knows only, say, two or three members. It becomes almost impossible if one only knows a single member. This problem is substantial because even for large databases, such as the genealogy databases in AUTOTYP (N = 2,680; Nichols & Bickel 2009) or in the *World Atlas of Language Structures* (N = 2,471; Haspelmath et al. 2005), about half of the stocks are represented by only one member (and this is so regardless of which of the two taxonomies is applied and even when one excludes, as I do here, creoles and sign languages, which could all be analyzed as single-member families in their own right since they represent the birth of new families).

The problem has both epistemological and statistical consequences. Epistemologically, the problem has the effect that many variables of typological interest cannot really be investigated when these variables happen to be best represented in less-well documented or isolated families. Statistically, the problem is one of power in detecting signals: any limitation to large families severely reduces the size of datasets, and this has the effect that statistical tests do no longer have enough power to detect signals.

In some sense, one could say that these are just the real limits on what one can possibly know about diachrony and the historical forces shaping typological distributions. However, to the extent that one can trust the results from examining biases in large families, it is possible to extrapolate from large to small families — i.e., for concreteness' sake, from families with at least 5 members (cf. Section 3) to families with less than 5 members. This requires two assumptions:

(4) a. *Normal Diachrony Assumption:*
The members of small families are the sole survivors of larger families.

b. *Uniform Development Assumption:*
Unknown families are subject to the same developmental principles as known families.

The rationale behind the Normal Diachrony Assumption is the following: if we don't know what other languages a language or a small group of languages is related to, i.e. if we are dealing with an isolate or a small group, this is only an epistemological issue, not an ontological one (a point also emphasized by Maslova (2000b)). Ontologically, isolates and small groups are still members of larger families; it's only that we don't know them because they became extinguished, in most cases because their speakers shifted to other, unrelated languages. This assumption is motivated by the fact that just like any other language, isolates and small families

must come from somewhere, i.e. they are the result of normal diachronic transmission (in the sense of Thomason & Kaufman 1988). Obviously, the assumption does not hold for most creoles and sign languages because they did not arise from normal diachronic transmission, and it would indeed be incorrect to extrapolate insights from known large families to the genesis of creoles and sign languages. (As a result, creoles and sign languages provide a different window on universal pressure shaping language than languages with a normal diachrony behind them and research on this requires other methods than what I discuss here.)

The Uniform Development Assumption assumption is again based on the insight that the status of languages as isolates or members of small groups is an epistemological and not an ontological fact: the extent to which we *know* genealogical relationship of a group has no principled consequences on the kinds of diachronic developments that the group went through. For example, just because we don't know any sister languages of Basque does not mean that the kind of diachronic processes that resulted in modern Basque was radically different from the kind of processes that resulted in the development of Hindi from Proto-Indo-European: we expect the same kind of complex mix of spontaneous (random) change, contact effects and universal pressure.

Taken together, the two assumptions in (4) allow us to extrapolate from large to small families. For this, we first compute the proportion of biased vs. diverse families in a survey of large families and then use this proportion as an estimate of the extent to which small families are biased. For example, in the survey of A-before-P orders in Section 3 we found that 69% large families are biased in some way and 31% are diverse. Based on (4), we can now make the assumption that the extent to which families diversify their ordering of A and P arguments, and, conversely, the extent to which families keep whatever order they have, does not only hold for large families but also for small families. In other words, we assume that our extremely reduced knowledge of Basque's ancestry has no consequences on how Basque developed (viz. by normal diachronic transmission, as per 4a) and to what extent the language was affected by universal pressure in language change (as per 4b). Therefore, we assume that about 70% of small families are the sole survivors of large families with a bias and about 30% of small families are the sole survivors of large families without a bias, i.e. to be diverse.

There is one probabilistic detail that we need to take care of before proceeding further, however: if we happen to find 100% large families to be biased (in whatever direction), it would not be correct to estimate a probability of 1 that small families are biased as well, i.e. there cannot be absolute certainty that all small families are biased and it is always possible that they represent larger diverse families. A well-established way of avoiding this is by estimating probabilities using Laplace's Rule of Succession: if k out of n large families are biased, we estimate the probability of small families to be biased as $\frac{k+1}{n+2}$. In our example of A-before-P orders, this would be $\frac{41+1}{59+2} = .689$. This is very close to the estimate based on the raw proportions (.695) but for smaller samples, the difference can be more substantial: if we had observed 10 out 10 families, the estimate would not be $\pi(\text{biased}) = 1$, but $\pi(\text{biased}) = \frac{11}{12} = .92$. The key idea behind the formula is this: the *a priori* assumption that families can in principle be either biased or diverse is equivalent to having observed one biased and one diverse family, and these 'as if'

observations are added to the observed frequencies.[7]

Using the estimated probability of being biased, we then randomly declare a corresponding proportion of small families to be the solve survivors of biased families.[8] In our example of A-before-P ordering, we would declare a random selection of 69% small families to be biased. The remaining small families (31%) are declared to be the sole survivors of diverse larger families. However, by virtue of being statistical estimates, biases are gradual and allow deviations: for example, one of the large families in the survey, Austronesian, is significantly biased towards A-before-P ordering. Despite this bias, 29 out of the 145 (20%) representatives of the family in the database deviate from this, partly by having VPA ("VOS") order (such as Kiribatese), partly by having variable word order (such as Acehnese). When assuming that a small family is the sole survivor of a biased family, the question therefore arises to what extent the small group we know represents the overall bias of the family, or deviates from this trend, just like the 20% of Austronesian languages that deviate from the overall trend in Austronesian.

In response to this, we first estimate the probability that a small group represents the family bias from the extent to which the bias is found in large families. As noted, in Austronesian this extent is .80; in other large families (e.g. Dravidian) it is 1. Large families can of course be biased in the opposite direction. In our example, we observed this for Algic and Iroquoian, and here the bias (against A-before-P) is in each case complete (i.e. 1) in the database. Taken all these bias estimates together suggests that, on average, if a large family is biased on the argument order variable in whichever direction (A-before-P or the opposite), it is so biased to 95.5%; and there are on average 4.5% deviates inside the family. (Austronesian, with as many as 20% deviates, is therefore quite exceptional.) When estimating the probability having a bias vs. being diverse, we corrected these estimates by the Laplace Rule of Succession because *a priori* it is always possible for families to be biased to some degree or to be diverse. For estimating the deviation probability, however, I suggest to rely on the bare proportions, i.e. if all biased families are completely biased, with no deviations, I suggest to assume a general deviation probability of 0. The reason is as follows. Postulating deviations is the same as postulating instances of language change (unlike postulating a bias, which may or may not imply language change, depending on the extent of the bias). Now, from general parsimony constraints on historical linguistics (Occams' Razor), it follows that one postulates language change only in the presence of positive evidence. Therefore, a priori – i.e. unless there is any evidence to the contrary – we assume that an isolated language or small language group represents its ancestors faithfully, with no change, no deviation.

Given these considerations, we can estimate the probability to which the members of a what we estimate is a biased small family represent indeed the family bias (here, 95.5%) and the probability to which members are likely to be deviating exceptions (here, 4.5%). Based on this, we randomly declare some proportion of the estimated biased families to be observed with representative members and some proportion to be observed with deviating members. In those small families where members are estimated to represent their family bias, we declare the

[7] It is a matter of further research to establish whether $\frac{1}{2}$ is indeed an appropriate parameter value of the *a priori* bias probability here.

[8] Technically, this is done via a randomly generated binomial distribution with the estimated bias probability.

family to biased towards whatever happens to be its sole type or what appears to be its most likely type given the general bias estimate and the frequency distribution within the family. For example, a small family will be declared to be biased towards A-before-P order if all or most of the small group have A-before-P order; if there is a tie (e.g. two languages with A-before-P and two languages with other orders), we randomly pick one as representative. In the small families where members are estimated to represent deviating exceptions, we declare them as survivors of a family that had a bias in an alternative direction (randomly chosen but weighted by the probability of directions given by the general bias estimate and the frequency distribution within the family). For example, if all or most languages in the small family have A-before-P order (or if indeed there is only a single language and it happens to have A-before-P order), we estimate that these languages come from a larger family with the opposite bias (i.e. no A-before-P order); if there is a tie, we again randomly select one of them.

In the overall extrapolation process there are three situation where we make random selections: first, when declaring a proportion of small families to be the sole survivor of diverse vs. biased families; second, when breaking ties for determining what kind of bias a small biased family represents; and third, when assigning an alternative type to those small biased families that we estimate as representing deviating exceptions within larger families. These random choices induce statistical error but because the error is random, it can be assumed to be normally distributed. Therefore, we can perform the extrapolations with all random selections many times (say, 2000 or 10,000 times) and then compute the mean of all extrapolation results.

For example, a single extrapolation might suggest 117 small families with an A-before-P bias, 29 with the opposite bias and 69 to be diverse; the next extrapolation might suggest 116 cases of A-before-P bias, 34 opposite biases and 65 diverse families etc. If we take the mean of these frequencies over 2000 extrapolations, we arrive at estimated frequencies of 115.07 A-before-P bias, 33.09 opposite bias and 66.83 diverse. Added to the estimates from large families, this results in an overall estimate of 154.07 A-before-P biases, 35.09 opposite biases and 84.83 diverse. This confirms the result from Section 3 that there is a significant trend for families to be biased towards A-before-P order as against P-before-A or free orders (exact binomial test, $p < .001$, $\hat{\pi} = .82$). [9]

In Section 2 I defined the strength of the universal pressure by the proportion of biased as opposed to diverse families. Since when extrapolating to small families, we use this proportion for estimating to what extent small families are the sole survivors of families with a bias, we can no longer re-compute this proportion from the extrapolation results. In other words, as far as I can see, estimates of the strength of universals can only be taken from large families. The estimate of the strength can therefore be defined as the Laplace estimator of biases discussed above, i.e.

$$s = \frac{k+1}{n+2}, \tag{5}$$

where k is the number of biased families out of a total of n families. Note that because of this equality, extrapolations will not be of help when testing for what I called Scenario **B** (non-

[9] A ready-to-use function for computing family biases in this way is available in an `R` package written by Taras Zakharko at `http://www.uni-leipzig.de/~autotyp/familybias.R`.

directional family bias) in Section 2: the proportion of biased families is by definition the same before and after the extrapolation. What can usefully be done, however, is to examine whether the bias is still undirected after extrapolation to small families.

Applying the extrapolation method to the non-directional family bias of predicative adjective encoding (cf. Section 3) suggests that this is the case. In Section 3 we found no statistical preference for any type. After extrapolation, we can estimate 34.63 families to be biased towards verbal, 30.48 towards nonverbal and 25.69 towards mixed encoding. These frequency estimates are still not significantly different from what one would expect under a uniform (⅓ each) distribution (χ^2 = 1.34, p = .54). This suggests that the encoding of property terms in predicative function is a genealogically stable property, not subject to any known universal or large-scale areal pressure.

In the two examples reviewed so far, the results of significance tests were not affected by the number of datapoints before vs. after extrapolation to isolates and small families. However, this can be quite different because (a) isolates and small families may bring in critical evidence and (b) because a statistical test may only be able to detect a signal if the dataset reaches a certain minimal size. This can be exemplified by examining the distribution of (some kind of) gender in independent pronouns, based on Siewierska's (2005) WALS dataset (N = 381, after adding a few data to which I had easy access in order to increase the number of stocks with more than five members[10]). Without extrapolations, the data from large families suggests non-directional family bias: of 17 large families, 12 have a bias; of these, 4 are biased towards having gender and 8 against. A proportion of 12 families out of 17 to be biased is borderline significant under a binomial test (p = .07, $\hat{\pi}$ = .71). But the 4:8 ratio in the direction of the bias does not suggest a significant preference (p = .194, $\hat{\pi}$ = .66), and so the data seem to suggest Scenario **B** from Section 2: daughter languages seem to maintain whatever the proto-language was like: if it had gender, gender is preserved; if it didn't have gender, it doesn't develop it. However, the absence of a statistical signal could also just reflect the fact that the total number of large families is very small (N = 8) and statistical tests don't have enough power to detect trends. In addition, there could be a possible direction in the bias specifically among small families and isolates. To find out, we can use the proportion of biases among large families (12 out of 17) and extrapolate to small families and isolates.

The mean extrapolations suggests that 42.24 (24%) families are biased towards and 80.55 (45%) against having gender distinctions in independent pronouns. This difference matches the 4:8 ratio among the large families, but the larger number now allows detecting a statistically significant signal (p < .001, $\hat{\pi}$ = .66). The result seems to suggest a universal bias against gender in pronouns.

However, in this case the extrapolations are based on only 16 families, and the result should not be taken as establishing a worldwide trend against gender. The only way to put the results on firmer grounds is to develop databases that collect more datapoints per family and thereby pursue a data-collection strategy that is the exact opposite of how most typologists have collected data in the past.

At any rate, as tentative as they are, the results on pronominal gender fit with Nichols's

[10] I added Tobelo, Galela, and Somali as languages with pronominal gender.

(1992, 2003) hypothesis that gender in general is disfavored universally. It does not seem to be particularly prone to inheritance unless there is support from neighboring families that have gender (like in Europe or Africa). Support from neighboring languages is an issue of areal conditions determining family biases, which is one of the topics in the following section.

6 Extending the Family Bias approach to multivariate distributions

In the preceding I have limited my attention to the distribution of a single variable. But the distribution of one variable may be conditioned by other variables, for example other structural variables (such as word order) or areal or social variables (such as Sprachbund membership or the presence of certain kinship systems).[11] I subsume all kinds of conditional effects under the term 'conditional pressure'.

Just like in univariate designs, conditional pressure can be strong or weak, and the lower bounds of this strength can be estimated by the proportion of biased vs. diverse families. Unlike in univariate designs, however, these strengths need not be uniform but can differ across conditions. For structural factors, pressure strength can be expected to be uniform across conditions in the case of bi-directional universals. An example is the classical hypothesis that OV structures favor postpositions and, conversely, that VO structures favor prepositions (Greenberg 1963, Dryer 1992). In the case of uni-directional universals (e.g. post-nominal relative clauses under non-verb-final word order conditions, but no trend towards pre-nominal relative clauses under verb-final conditions), one expects strong pressure in one condition (here, under the non-verb-final condition) but under the other condition, there can be many diverse families, or families can be biased in random ways. The universals is supported as long as the trend towards biases is stronger under one than under the other condition.

For areal factors, such differences are in fact expected: when comparing the distribution of features inside vs. outside an area, one expects strong pressure towards some feature *F* (less diversity) inside the area but only weak pressure against *F* (more diversity) outside the area. Areal diffusion leads to the widespread adoption of *F*, resulting in an increased frequency of *F*. This is in contrast to the world outside the area, where nothing is suspected to affect the distribution of *F*: it can tend to be diverse within families or families can be biased in random ways. All that matters for areality is that there is significantly higher proportion of families with an *F*-bias inside than outside the area.

Testing bivariate hypotheses like these is complicated by the fact that the relevant condition may not hold for entire families: for example, stocks like Indo-European or Sino-Tibetan contain both VO and OV branches. A solution to this problem comes from the fact that the Family Bias Theory makes no assumptions about the taxonomic level or time-depth at which biases can be found; it is not even required that biases are always found on the same taxonomic level across families (cf. Section 2). This suggests that family biases can be estimated in whatever is the highest taxonomic level at which subgroups are not split with regard to the relevant condition. In Indo-European for example, one can estimate biases within OV and VO branches.

[11] or many such variables together and interacting with each other. Here I concentrate on simple cases. For a discussion of interactions between conditioning variables, see Bickel (2008a) and Cysouw (2010a).

There is one further complication, though: given the often sketchy knowledge that is available on subgrouping, it is often impossible to find plausible subgroups; or, even if the taxonomy is well established, subgroups may be diverse with regard to some condition of interest. In both these cases, I propose to posit 'pseudo-groups', based on the difference in the relevant conditions, e.g. an OV pseudo-group vs. a VO pseudo-group. Importantly, these pseudo-groups are posited solely for the purposes of testing whether differences in the condition have an effect on family biases. They are not evidence for real subgroups because changes in typological properties can be due to factors that are entirely different from the kind of arbitrary and idiosyncratic innovations that define genealogical trees. However, since some change must have split the family, it is a legitimate isogloss for testing purposes: under the Family Bias Theory, the question is only whether the isogloss is associated with different responses to such an extent that the pseudo-groups are now biased in a predictable direction.

In the following I first exemplify this approach with hypotheses on structural and then on areal factors.

6.1 Example 1: relative clause position and word order

The first example concerns the hypothesis that the odds for relative clauses to be post-nominal are higher under non-verb-final than under verb-final conditions (Greenberg 1963, Dryer 1992, Hawkins 1994, among others). To examine this hypothesis, I combined Dryer's (2005a) WALS dataset on relative clause position with his dataset on dominant main clause verb positions, excluding languages with flexible orders (Dryer 2005b), but adding more critical data on Sinitic (Yue 2003). Both small and large families can be homogenous or split on the relevant condition, i.e. can contain both non-verb-final and verb-final languages. In the dataset (N = 513 languages), there are 22 large stocks (i.e. with at least 5 members) and 96 small families. Of the 22 stocks, 14 (or 64%) are homogeneously verb-final or non-verb-final. In 4 (18%) stocks (Indo-European, Sino-Tibetan, Cushitic, and Austroasiatic), homogenous branches can be found at the major branch level.

In some cases, determining family biases at lower levels leaves small families or even single languages stranded as the sole representatives of their branch, e.g. Albanian and Greek in Indo-European, or Lolo-Burmese (with 3 representatives), two Naga languages and few others in Sino-Tibetan. In some cases, an entire branch ends up with single-member groups: the Western Oceanic group of Austronesian, for example, is represented in the database by Tolai and Tawala. Since the two languages differ in basic word order, they are assumed here to represent their own single-member groups.

In 3 of the 22 stocks (13%), homogenous groups can be found only by positing pseudo-subgroups. One example is Uto-Aztecan. While the Northern branch is homogeneously verb-final, and the Aztecan group of the Southern branch is homogeneously non-verb-final, the Sonoran group of the Southern branch is mixed. There are two non-verb-final and five verb-final languages and the distinction does not match any subgrouping represented in the AUTOTYP taxonomy assumed here (although it might of course fit other possible subgrouping hypotheses). In this case, I posit two pseudo-subgroups for computing family biases, a non-verb-final one and a verb-final one. Another example is found in the Bantu branch of Benue-Congo: all but one Bantu language in the database are non-verb-final. Here I posit a large non-verb-

	final	non-final	Sum
diverse	2	0	2
Rel-N	6	1	7
N-Rel	1	17	18
Sum	9	18	27

(a) large families only

	final	non-final	Sum
diverse	22.79	5.80	28.59
Rel-N	29.73	1.95	31.68
N-Rel	32.48	127.25	159.73
Sum	85.00	135.00	220.00

(b) with extrapolation to small families

Table 2: Family biases in relative clause position dependent on main clause verb position

final pseudo-group and a small verb-final pseudo-group which is represented only by a single language in the database (viz. Tunen: Mous 1997). The third case where pseudo-groups are necessary is Arawakan. This stock is represented in the database with only single representatives from each branch, with five non-verb-final and one verb-final branch. Here I assume a non-verb-final pseudo-group with five members and a small single-member group.[12]

With this, we arrive at a total of 27 large families, including 3 pseudo-groups. Of the 27 families, 14 (52%) are at the stock level, 9 (33%) at the highest ('major') branch level, and 4 (15%) at lower levels. Table 2a cross-tabulates family biases against main clause verb order. The hypothesis is that biases towards N-Rel (Noun-Relative Clause) sequences are much more likely in non-verb-final than in verb-final families. This can be tested with a Fisher Exact Test comparing the odds for families with N-Rel biases against families with the opposite bias under verb-final vs. non-verb-final conditions. The result suggests a significant and strong effect (one-sided $p <$ .001, estimated odds ratio[13] $\hat{\theta} = 64.64$). This is obviously caused by the fact that only one non-verb-final family in the database (viz. Sinitic) is biased towards pre-nominal relative clauses. The strength of the universal can be estimated by the Laplace estimator of the probability of biases (cf. 5), suggesting a pressure of $\hat{s} = .95$ under the critical condition of non-verb-final order, i.e. a fairly strong universal.

Under the other condition, verb-final order, the bias probability is $\hat{s} = .73$. This suggests that under verb-final conditions, relative clause position is genealogically stable (Scenario **B** in Section 2) or, alternatively, that there is a universal trend favoring pre-nominal clauses (Scenario **A**). The 6:1 ratio in Table 2a is suggestive of a directional trend, but the counts are small and exclude data from small families and isolates.

In response to this, I performed extrapolations following the same procedure as described in Section 5, separately for each condition. Using the bias probability estimates of .95 for non-final and .73 for final word order, this results in the mean estimates summarized in Table 2b.

[12] The single verb-final Arawakan language in Dryer's (2005b) database is Tariana, but this language would in fact seem to be more accurately coded as lacking a dominant order (Aikhenvald 2003). On either analysis, Arawakan requires pseudo-groups until possible subgroupings are robustly established.

[13] Although not commonly used in typology, the odds ratio (θ) is a standard and useful statistic to compare proportions across conditions. It is defined as the ratio between the odds, and so an odds ratio of about 66 means that the odds for biases towards post-nominal relative clauses are 66 times higher in non-verb-final than in verb-final families.

The extrapolations confirm the first finding from the large families: there is a significant trend for families to be biased towards post-nominal relative clauses under non-verb-final conditions (Fisher Exact test, p < .001, $\hat{\theta}$ = 57.98). The mean estimated number of non-verb-final families with pre-nominal relative clauses is 1.95. This figure results from the fact apart from Sinitic, in 1895 out of 2000 extrapolations (i.e. in 95%), an additional non-verb-final language was estimated to represent a family with an original bias towards pre-nominal relative clauses. This language is the Sino-Tibetan language Bai (Wiersma 2003), which has SVO main clauses and pre-nominal relative clauses. The exact position of Bai within Sino-Tibetan is controversial, and Nichols & Bickel's (2009) taxonomy treats Bai as a stock-level isolate. It is possible that Bai comes from a branch that originally had post-nominal relative clauses and changed to pre-nominal order under Sinitic influence, i.e. that Bai comes from a diverse branch. It is also possible, however, that Bai inherited pre-nominal relative clauses from one of its ancestors (which might have changed to pre-nominal order earlier, again possibly under Sinitic influence or even identity with proto-Sinitic). Without further reconstruction and detailed comparative work, it is impossible to decide between these scenarios. All that we know for good is that Bai now has pre-nominal relative clauses and that, worldwide, the position of relative clauses in non-verb-final languages is relatively stable over time (estimated at $\hat{s}$ = .95). This favors a scenario whereby Bai inherited its ordering principles from its branch ancestor and does not reflect recent change under Sinitic influence. In the extrapolations, this high bias estimate of .95 results in Bai being taken to reflect a bias towards pre-nominal relative clauses in 95% of the 2000 extrapolations, pushing the mean number up to 1.95.

The second finding from the large families was that there is a possible preference for pre-nominal relative clauses under verb-final conditions. Table 2a suggests odds of 6:1 for this. But this is not confirmed by the extrapolations in Table 2b, where the odds (29.73:32.27 = .92) go in a different direction but are not significantly different from 1 anyway (p = .48, $\hat{\pi}$ = .65). This makes it likely that there is no directional bias and that instead the bias strength of $\hat{s}$ = .73 reflects a fair degree of genealogical stability (cf. Scenario **B** in Section 2).

6.2 Example 2: hotbeds of pronominal gender

At the end of Section 5 we observed tentative evidence for universal pressure against pronominal gender. However, as Nichols (1992, 2003) notes for gender in general, pronominal gender tends to be better retained in families when they cluster together with similar families in 'hotbeds': while the phenomenon does not appear to spread easily, its retention seems to be favored in specific regions.

In order to test this hypothesis, I classified the data from Siewierska (2005) into three hotbeds based on Nichols's (1992) suggestions: Africa (including adjacent Semitic languages), (Western) Europe (up to a line from the Carpathians following the Wisła to the Baltic see, cf. Nichols & Bickel 2009) and the Sahul area (including near islands up to the Wallace line and collapsing the strata postulated by Nichols 1997b; see cf. Map 1). I then computed family biases within and outside the hotbeds.

The dataset contains 17 large families (with at least five members), 35 small families, and 125 isolates. Most families are located completely within or outside hotbeds, but 3 of the 52 (6%) families are split: Indo-European, Uralic, and Austronesian. Within Indo-European, non-

Map 1: The distribution of pronominal gender across hotbeds (Siewierska 2005, with some additions in South America and Papua New Guinea). ●/○ = Africa; ■/□ = (Western) Europe; ◆/◇ = Sahul; ▲/△ = rest of the world; filled symbols denote presence, empty symbols absence of pronominal gender

split taxa can be found at the major branch level except for Balto-Slavic which according to Nichols & Bickel's (2009) narrow definition of Europe splits into subbranches within (West Slavic) and subbranches (East Slavic and Baltic) outside the European hotbed. The same is true of Uralic, where non-split taxa can only be found at the lowest taxonomic levels since even Finno-Ugric is split by the narrow definition (leaving Hungarian inside the European hotbed and Finnish outside). The situation is again similar in Austronesian where the Sahul hotbed boundary crosscuts the Oceanic and Central Malayo-Polynesian subgroups. As a result, non-split groups can only be found at relatively shallow taxonomic levels, each with small numbers of members (below 5). The rest of Malayo-Polynesian (the 'Western Malayo-Polynesian non-clade' of Nichols & Bickel 2009) is again split and for lack of established subgrouping, it is divided here for statistical purposes into a large pseudo-group (N = 8) outside and smaller pseudo-group (N = 4) inside the Sahul hotbed.

The splits leave a total of 18 large unsplit groups (with more than five members), tabulated in Table 3a. This is a small number to base statistical estimates on. For the Laplace estimator (cf. 5) this means lost of precision and more random guessing on the extent to which small families represent biases. To some extent this is compensated by the resampling strategy described in Section 5, but the results must clearly be taken as preliminary. On the basis of Table 3a the bias estimator is $\hat{s}$ = .70 inside and $\hat{s}$ = .58 outside the hotbeds: as expected, it is a bit more likely for families to be biased (in some direction) inside than outside the hotbeds. The bias estimators result in a mean extrapolation to small families as given in Table 3b. An analysis of the extrapolations shows that the odds for biases towards gender are $\hat{\theta}$ = 2.3 times higher inside than outside the hotbeds, which is significant under a Fisher Exact Test (p = .029). This supports the hypothesis that gender is better preserved in families when they cluster together in hotbeds.

	inside	outside	Sum
diverse	2	4	6
with gender	4	0	4
without gender	2	6	8
Sum	8	10	18

(a) large families only

	inside	outside	Sum
diverse	29.03	47.73	76.76
with gender	29.10	51.08	45.29
without gender	39.87	16.19	90.95
Sum	98.00	115.00	213.00

(b) with extrapolation to small families

Table 3: Family biases in pronominal gender inside vs. outside hotbeds

7 Discussion

The family bias estimates reported here are preliminary and clearly need far more denser sampling of families. This is a sampling strategy that is the exact opposite of what has been recommended in the past, where typologists have emphasized that samples should avoid picking many representatives from the same families, i.e. that they should be genealogically balanced. It is instructive to compare the results reported here to the conclusions that one might draw on the basis of genealogically balanced sampling.

Since Dryer (1989), the standard in the field has been to create samples in which each family contributes one single datapoint. When families are diverse, e.g. some members have gender and others don't, they are sometimes counted as contributing several datapoints (and there are more or less refined methods for dealing with this, cf. Bickel 2008b). The key point of the method, however, is that families are always treated statistically in the same way as isolates and that any trends or biases within families is ignored. The result of such a per-family count (using Bickel's (2008b) algorithm) is given in Table 4. Unlike in Table 3b, the odds ratio of this table is not near significance under a Fisher Exact test ($\hat{\theta} = 1.45$, $p = .21$), i.e. there is no evidence for a higher chance of hotbed languages to have pronominal gender.

	inside	outside	Sum
with gender	39	42	81
without gender	58	91	149
Sum	97	133	230

Table 4: The distribution of pronominal gender inside vs. outside hotbeds in a genealogically balanced database

The difference in the results is not one of statistical power since the sample sizes are comparable. The difference is a matter of methodological principle. Genealogically balanced sampling makes the implicit assumption that if a feature is shared by the daughter languages of a family, this can only reflect faithful replication, with no other motivation or cause than sheer inertia: a feature is preserved just because the parent language had it and speakers are conservative. Therefore, if one wants to test for factors that might influence the feature, such as location in-

side a hotbed or some structural condition (e.g. word order), one should not count all languages inside the same family as independent datapoints but instead reduce the data to genealogically independent datapoints, i.e. a genealogically balanced dataset.

However, there is no reason to assume that retention must be free of conditions or causes: in fact, retention can and often is favored by universal preferences or areal factors. This becomes particularly clear in hypotheses that are specifically about the conditions under which features are best retained (most stable) — such as Nichols's (2003) hypothesis on gender that is tested here: what is at stake is whether families tend to retain pronominal gender more often inside rather than outside hotbeds. Because retention of a feature in a family necessarily leads to a bias towards that feature in the synchronic daughter languages, this hypothesis directly predicts that we find more families biased towards gender inside than outside hotbeds. The results in Table 3b confirm this.

Genealogically balanced sampling, by contrast, removes all data we have on inheritance patterns within families and therefore makes it impossible to test the hypothesis. As a result, the data in Table 4 display some aspects of the synchronic distribution of gender, but it does not allow any inference on the processes that lead to this distribution. But this is in conflict with the very nature of typological hypothesis: synchronic distributions (except perhaps for those of creole and sign languages) must come from somewhere, and the only way a typological factor can play a role in these distributions is by affecting the way languages change over time. Since they are hypothesis on diachronic changes, the only possible way to test them is to try and estimate the extent to which changes lead to systematic biases across families.

8 Conclusions

In this chapter I proposed a theory that links types of family distributions to specific historical scenarios. The key evidence for the theory is that the number of typological changes in known families is far higher than what would be expected if typological distributions had persisted over deep time, going back to the origins of the human language faculty. This casts doubt on any attempt to use typological data for assessing the kind of language that the first representatives of our species spoke. At the same time, the theory proposed here suggests that typological distributions are systematically driven by the interaction of high-fidelity replication with various kinds of external pressure, such as universal principles and areal diffusion trends.

Therefore, distributional biases in families do not allow a direct and generalizing estimation of specifically genealogical stability (*pace* Parkvall 2008, Wichmann & Holman 2009 or Bakker et al. 2009). Any such estimation needs to factor in the possible effects of external pressure, and this means that any progress in estimating stability indices depends on our knowledge of such pressure, including the effects of universals. Rather than leading to sweeping hypotheses like 'pronominal gender is stable', the analyses presented above suggest that pronominal gender is significantly more stable inside than outside hotbeds, but that in both situations there is in fact a trend for families not to develop gender in the first place. Similarly, instead of determining whether the position of relative clauses is generally stable or unstable, we found that this depends on word order conditions: under verb-final conditions, relative clause position is genealogically fairly stable, i.e. it seems to just follow whatever the ancestor language had. But

under non-verb-final conditions, there is strong universal pressure for biasing families towards post-nominal position.

This confirms Nichols's (2003) insight that the overall stability of a typological property "is a matter of competing forces" and typically results from the combined effects of faithful replication and external pressure. A full understanding of language change and typological distributions must simultaneously engage in research on universals, language contact effects, and the extent to which patterns are faithfully replicated. There is no shortcut.

These findings also have a practical consequence: because all estimates of inheritance and external pressure, including any extrapolation to isolates, are based on distributions in known families, typological databases need to sample families as densely as possible. This suggests a radical move away from the classical 'one-language-per-family' sampling strategy that has dominated database development in the past.

References

Aikhenvald, Alexandra Y., 2003. *A grammar of Tariana.* Cambridge: Cambridge University Press.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, & Eric W. Holman, 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typology* 13, 169 – 181.

Bickel, Balthasar, 2007. Typology in the 21st century: major current developments. *Linguistic Typology* 11, 239 – 251.

Bickel, Balthasar, 2008a. A general method for the statistical evaluation of typological distributions. Ms. University of Leipzig, [`http://www.uni-leipzig.de/~bickel/research/papers/testing_universals_bickel2008.pdf`].

Bickel, Balthasar, 2008b. A refined sampling procedure for genealogical control. *Language Typology and Universals* 61, 221–233.

Bickel, Balthasar & Johanna Nichols, 2005. Exponence of selected inflectional formatives. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 90 – 93. 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 Hawkins, John A. (ed.) *Explaining language universals*, 350 – 379. Oxford: Blackwell.

Cysouw, Michael, 2010a. Dealing with diversity: towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14, 253–286.

Cysouw, Michael, 2010b. On the probability distribution of typological frequencies. In Ebert, Christian, Gerhard Jäger, & Jens Michaelis (eds.) *The Mathematics of Language*, 29 – 35. Springer.

Dahl, Östen & Viveka Velupillai, 2005. Tense and aspect. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 266 – 282. Oxford: Oxford University Press.

Daniel, Michael, 2005. Plurality in Independent Personal Pronouns. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 146–149. Oxford: Oxford University Press.

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

Dryer, Matthew S., 1992. The Greenbergian word order correlations. *Language* 68, 81 – 138.

Dryer, Matthew S., 2005a. Order of relative clause and noun. In Haspelmath, Martin, 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 Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 330 – 341. Oxford University Press.

Dryer, Matthew S., 2005c. Polar questions. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 470–473. Oxford: Oxford University Press.

Dryer, Matthew S., 2005d. Position of case affixes. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 210 – 213. Oxford: Oxford University Press.

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*, 73 – 113. Cambridge,

Mass.: MIT Press.

Greenberg, Joseph H., 1995. The diachronic typological approach to language. In Shibatani, Masayoshi & 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 Hawkins, John A. (ed.) *Explaining language universals*, 321 – 349. Oxford: Blackwell.

Haspelmath, Martin, 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18, 180 – 205.

Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.), 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

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

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

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

Maslova, Elena & Tatiana Nikitina, 2007. Stochastic universals and dynamics of cross-linguistic distributions: the case of alignment types. Ms. Stanford University, `http://anothersumma.net/Publications/Ergativity.pdf`.

McBrearty, Sally & Alison S. Brooks, 2000. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *Journal of Human Evolution* 39, 453 – 563.

Mous, Maarten, 1997. The position of the object in Tunen. In Déchaine, Rose-Marie & Victor Manfredi (eds.) *Object Positions in Benue-Kwa*, 123–137. The Hague: Holland Academic Graphics.

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

Nichols, Johanna, 1997a. Modeling ancient population structures and population movement in linguistics and archeology. *Annual Review of Anthropology* 26, 359 – 384.

Nichols, Johanna, 1997b. Sprung from two common sources: Sahul as a linguistic area. In McConvell, Patrick (ed.) *Archeology and linguistics: global perspectives on Ancient Australia*. Melbourne.

Nichols, Johanna, 2002. Monogenesis or polygenesis? Typological perspectives on language origins. Plenary lecture at the Annual Meeting of the Linguistic Society of America, January 3, 2002.

Nichols, Johanna, 2003. Diversity and stability in language. In Janda, Richard D. & Brian D. Joseph (eds.) *Handbook of Historical Linguistics*, 283 – 310. London: Blackwell.

Nichols, Johanna & Balthasar Bickel, 2009. The AUTOTYP genealogy and geography database: 2009 release. Electronic database, `http://www.uni-leipzig.de/~autotyp`.

Parkvall, Mikael, 2008. Which parts of language are most stable? *Language Typology and Universals* 61, 234 – 250.

R Development Core Team, 2011. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, `http://www.r-project.org`.

Siewierska, Anna, 2005. Gender distinctions in independent personal pronouns. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 182–185. Oxford: Oxford University Press.

Stassen, Leon, 2005. Predicative adjectives. In Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *The world atlas of language structures*, 478 –481. Oxford: Oxford University Press.

Thomason, Sarah Grey & Terrence Kaufman, 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Wichmann, Søren & Eric W. Holman, 2009. *Temporal stability for linguistic typological features*. Munich: LINCOM EUROPA.

Wiersma, Grace, 2003. Yunnan Bai. In Thurgood, Graham & Randy J. LaPolla (eds.) *The Sino-Tibetan languages*, 651 – 673. London: Routledge.

Yue, Anne O., 2003. Chinese dialects: grammar. In Thurgood, Graham & Randy J. LaPolla (eds.) *The Sino-Tibetan languages*, 84–123. London: Routledge.