A dynamic approach to the verification of distributional universals

ELENA MASLOVA

Abstract

Although distributional universals play a major role in typological studies, there are no reliable empirical criteria which would distinguish between genuine distributional universals and accidental statistical properties of the current language population. Progress in this domain has been hindered by the inadequacy of the "languages-as-trials" approach (implicitly adopted in current typological practice), which fails to assess random effects of "historical accidents" and to account for the dependency of language type on the properties of the ancestor language. This paper puts forward a model that takes into account both types of phenomena, thereby determining types of statistical evidence required for the empirical verification of distributional universals. The model is based on the notion of Markov process, which is applied to explore the effects of two stochastic processes which bring about typological distributions: the birth-and-death process in the language population and the process of type-shifts in the history of each language. The notion of a distributional universal is defined as the stationary distribution of a type-shift process, i.e., as the distribution that would in the long run be achieved by a set of independently developing languages.

Keywords: *birth-and-death process, distributional universal, Markov process, methodology, probability sampling, type-shift*

1. Introduction

The role of DISTRIBUTIONAL, or statistical, UNIVERSALS in modern linguistic typology is based upon two observations: (i) the languages of the world display a number of statistical tendencies which need to be accounted for (Comrie 1989: 19–20), and (ii) at least some of these tendencies appear to be linguistically motivated, which suggests that they are not "accidental", cf. Hawkins (1990: 96):

Linguistic Typology 4 (2000), 307–333

1430–0532/2000/004-0307

> [...] only those co-occurrences and language frequencies will ultimately be significant that CAN be motivated and explained by some theory, thereby distinguishing them from those that are accidental or attributable to historical forces of no relevance for linguistics, such as large armies!

Yet it is hardly necessary to argue that a THEORETICAL explanation of an observation cannot substitute for a proof of the EMPIRICAL validity of that observation. More specifically, no linguistic explanation of a distribution pattern can guarantee that this pattern is determined by linguistic principles and not by "historical forces".[1]

The problem is that there seem to be no criteria that would allow for an empirical distinction between genuine distributional universals and accidental statistical tendencies. This is known in typology as the problem of PROBABILITY SAMPLING: it turns out to be impossible to select a language sample which would be sufficiently large for statistical conclusions and at the same time meet the requirement of mutual independence between units, as implied by any statistical test (Dryer 1989: 262–263, Rijkhoff & Bakker 1998: 264–265). Croft concludes his critique of Dryer's (1989) method, which is intended to overcome this problem, by the following statement (1995: 91):

> This is an essentially inescapable problem, and can only be surmounted by obtaining evidence for typological explanations from other sources of data.

Rijkhoff & Bakker, in the most recent discussion of the problem, arrive at essentially the same conclusion, namely, "that there does not seem to be a real solution" (1998: 265). Yet if this is the case, then the notion of a distributional universal, however promising theoretically, simply lacks any empirical grounds and must be renounced; there are simply no rational reasons to believe that language frequencies can be linguistically significant. Implications of this conclusion for typological studies are more far-reaching than it might seem at first sight, since it pertains not only to "statistical" universals, but to ALL empirical universals: if a logically possible type is absent from the current language population, this can also be an "accident". Most importantly, the notion of distributional universal is explicitly or implicitly invoked by all empirically established INTERDEPENDENCIES between different parameters of linguistic variation, which constitute, as the editorial statement of this journal reads, "the essence of typology". Thus, the problem of the empirical verification of distributional universals is of vital importance for the very ability of linguistic typology to achieve its stated goals. If it is indeed insoluble in principle, the goals must be adjusted to the limitations of the method.

The present paper is intended to show that this is not the case, that is, there exists a theoretically justified way to solve this problem. The approach suggested here is based upon two cornerstones. First, the model of INDEPENDENT

TRIALS, which has proved inadequate for typological phenomena, is replaced by a more powerful model of MARKOV CHAINS, so as to capture the obvious fact that the current type of each language depends on the linguistic properties which this language (or its ancestor) happened to have at the previous stage of its history (cf. Greenberg 1978, 1995: 146–153). The essential inability of the model of independent trials to account for such dependencies is in fact the major obstacle that has hindered progress in this domain (Sections 2.2–2.3). Secondly, the potential impact of "historical forces" on distributions of language types is assessed by means of an analysis of the RANDOM BIRTH-AND-DEATH PROCESS in the history of the language population (Section 2.1). So called "historical accidents" are thereby shown not to represent an "inescapable" obstacle, because their distorting effects can be dealt with in a mathematically reliable way (Section 3). These two steps open a theoretical possibility to design reliable procedures for the empirical verification of distributional universals (Section 4).

2. Basic definitions and assumptions

2.1. *Language populations and A-distributions*

A LANGUAGE POPULATION is the set of all languages which exist in the world at some moment of time (cf. Nichols 1993: 2–3). It is assumed that, at any given moment of time, it is possible to identify units which constitute a language population, i.e., languages. The plausibility of this assumption, which in effect boils down to the validity of the very notion of "language", is outside the scope of the present paper. Suffice it to note that the claims to be argued for here are not sensitive to the criteria employed to distinguish languages from dialects and other types of language varieties, provided of course that these criteria are not biased in favor of one or another language type being investigated.

A typology $\mathbf{T} = \{T_i\}$ is a classification of languages according to any parameter of linguistic variation or a combination of such parameters. It is assumed that each member of any language population belongs to exactly one type T_i in any given typology. This condition is purely formal, since it can be fulfilled for any typology by introducing, if necessary, a special type for languages lacking the phenomenon under investigation and/or "mixed" or "intermediate" types. Any population is characterized by the distribution of its members over the types of a given typology, i.e., by frequencies of these types in the population. Such distributions will be referred to below as A-distributions (where "A" can be read as either "actual" or "accidental").

An A-distribution in the current language population can be estimated on the basis of a representative sample by means of the usual statistical techniques; an exemplary study of this sort is Tomlin (1986). However, it is implausible to interpret an A-distribution in terms of general linguistic principles (not to speak

of inferring such principles from A-distributions) without further evidence that this distribution is not an accidental property of this particular population observed at this particular time. The point is that the current A-distributions are not the only possible ones; in particular, there are good reasons to believe that earlier language populations displayed A-distributions essentially different from those observed at the present time (Dryer 1989).

There are two types of events which can, in principle, modify an A-distribution. First, a language can cease to exist (LANGUAGE DEATH) or a language variety can develop into a separate language (LANGUAGE BIRTH), thereby decreasing or increasing the total number of languages of a certain type. Secondly, a language can change from one type towards another type (TYPE SHIFT). The major difference between these event types with regard to distributional universals resides in their relation to the linguistic properties of the languages involved. The birth-and-death events are assumed to take place independently of linguistic properties, that is, the probabilities for a language to split p(B) and to die p(D) are viewed as attributes of the entire population, which do not vary depending on language type. In contrast with this, the probability of a type-shift saliently depends on the current linguistic properties of a language. In terms of Hawkins (1990: 96), the birth-and-death process represents "historical forces with no relevance for linguistics", whereas "linguistic forces" can only exhibit themselves through type-shift processes. Note that the general assumptions outlined above entail that events of both types are conceived of as instantaneous: if at any moment of time a language population comprises a set of identifiable members which belong to particular types, it means that any event which modifies either the population itself or an A-distribution in that population occurs between two subsequent moments of time. This apparent oversimplification can be compensated for by introducing an appropriate time scale; in the present paper, it will be convenient to take 100 years as the minimal unit of time.

Thus, the current A-distributions have been brought about by a combination of three heterogeneous factors. The first factor is what may be referred to as INITIAL A-DISTRIBUTIONS, i.e., the linguistic properties of the proto-language or the A-distributions in a proto-population. Further, the language population undergoes two types of stochastic (i.e., non-deterministic) processes, (i) a BIRTH-AND-DEATH process in the population and (ii) TYPE-SHIFT processes in the history of each individual language. However, for a discussion of language universals it will be reasonable to choose the "initial" moment of time t_0 in such a way as to ensure that languages existing at t_0 be at the same "level of evolution" as their current descendants (Comrie 1989: 8), so that all languages under consideration could be assumed to conform to the same universal principles. Then, the initial A-distribution for a given typology is the A-distribution that happened to exist at t_0. Following Comrie (1989: 9), I will

adopt ten thousand years ago as an appropriate estimate for t_0. As will be clear from the following discussion, this estimate is adopted only for the sake of convenience. Another choice of t_0 would lead to slightly different interpretations, but would not affect the major conclusions (Section 4.1).

It may seem that this model ignores areal phenomena, yet this is not the case. To begin with, the model does not involve any assumptions about the factors that cause each particular event: evidently, language contact is a possible determinant factor; yet a contact situation affects an A-distribution only if it occasions either a type-shift or a birth-and-death event (see also Section 2.2). On the other hand, as will be shown in Section 4.2, large-scale areal phenomena, which have been pointed out by Dryer (1989) as a significant source of "distorting effects" in language sampling, are straightforwardly accounted for by the model.

2.2. *Distributional universals and transition probabilities*

The notion of DISTRIBUTIONAL UNIVERSAL is commonly defined just by stipulating a necessary condition on language samples which might be used to establish such universals: it is required that all instances of linguistic types in a sample be mutually independent (Perkins 1989, Rijkhoff & Bakker 1998: 265). This requirement is based on a general principle of statistics, which states that the frequency of an event in a set of independent trials serves as a reliable estimate of the probability of that event, and hence implies that the distributional universal for a given typology $\mathbf{T} = \{T_i\}$ is conceived of as a set of probabilities $\mathbf{P} = \{p(T_i)\}$ for a language to be of type $T_i(\sum p(T_i) = 1)$. Languages are thereby construed as "trials", the possible outputs of each trial being types T_i of the typology. This "languages-as-trials" approach involves an important assumption which has scarcely been recognized as non-trivial, namely, that each typology **T** is associated with a UNIQUE universal distribution **P**, which determines the probabilities of possible outputs for each language-trial. The linguistic counterpart of this apparently formal prerequisite is that the probabilities constituting a distributional universal are determined solely by linguistic preferences.

The problem is that the assumption of uniqueness, in its plain form inherent in the languages-as-trials approach, is false. This will be fairly clear if we imagine a language population whose history began with a large set of absolutely independent languages. For the sake of simplicity, it can be assumed that this population does not undergo the birth-and-death process, so that the problem of mutual dependencies between linguistic properties of languages is ruled out. Nonetheless, the A-distribution in this population will change with time by virtue of type-shift processes, so that the population may display significantly different A-distributions at different stages of its history. Furthermore, since

type-shift events are obviously not independent of the current linguistic properties of a language, these A-distributions will be to some extent determined by the A-distribution which happened to exist at the start. Since the languages-trials in this imaginary world are always independent, the frequency of each type T_i at any moment of time would reflect some probability $p(T_i)$, yet this probability would vary with time, which entails that there exist multiple distributions **P** associated with the same typology. What follows is that the notion of distributional universal can only be appropriately defined if it is possible to single out one of these distributions in such a way as to maintain the underlying linguistic intuition, that is, if there exists one and only one distribution **P** which can be conceived of as a manifestation of linguistic preferences.

One essential condition on this distinguished distribution is obvious: it must be INDEPENDENT of the initial type of a language. In other words, for the notion of distributional universal to make sense, it is required that, after some (albeit possibly long) period of time, the probability for a language to belong to a certain type does not depend on the type which this language (or its ancestor) happened to have at t_0. As pointed out by Greenberg (1995: 146–147), this requirement can only be satisfied if the types of a typology are strongly connected by the type-shift process, that is, if there is a diachronic path from any type to any other type, possibly through some intermediate stages. Indeed, if there is no path leading from some type T_i to another type T_j, the probability $p(T_j)$ will always be zero for languages whose initial type happened to be T_i, whereas for some other initial types this probability will be non-zero. It follows that the dependency of $p(T_j)$ on the initial type will last forever. Thus, the notion of distributional universal can only be defined for strongly connected typologies.

Another condition directly follows from the requirement of uniqueness: once the independent distribution is achieved, it must remain STABLE. This is only possible if there exists a sort of equilibrium between the synchronic distribution and the probabilities of type-shifts, so that, for each type T_i, the total number of languages changing FROM T_i towards other types within any time interval is roughly equal to the total number of languages changing TOWARDS T_i within the same time interval. This condition gives the following set of equations:

$$(1) \qquad p(T_i) = \sum_{k=1,\ldots,n} p(T_k)p(T_k \rightarrow T_i),$$
$$\sum_{i=1,\ldots,n} p(T_k \rightarrow T_i) = 1,$$

where n is the number of types in the typology, the TRANSITION PROBABILITY $p(T_k \rightarrow T_i)$ is the probability that a language of type T_k will be found in state T_i after a small time interval[2] (for $k = i$, it is the probability to retain the type, otherwise, the probability of a type-shift). This is the so-called

STATIONARY DISTRIBUTION, which is uniquely determined by the underlying type-shift process.[3] It is important to stress that equations (1) define the probabilities for a SINGLE language to be found in different type-states. Therefore, it licenses both a diachronic and a synchronic interpretation: the probability $p(T_i)$ can be thought of either as the probability that a language will be found in state T_i at a randomly selected moment of its HISTORY or as the probability for a randomly selected member of a large language POPULATION to be in this state (cf. Hawkins 1983: 256).

According to the criterion of uniqueness, the stationary distribution is the only distribution **P** which may, in principle, count as the distributional universal for a given typology. The question is whether the stationary distribution can be plausibly assumed to be motivated by linguistic preferences. Since it is determined by transition probabilities, this question boils down to the well-known problem of motivations for language change: in effect, a definition of a distributional universal in terms of the stationary distribution amounts to the assumption that transition probabilities reflect linguistic preferences (cf. Hawkins 1990), whereas "external" causes of language change (most importantly, language contact) can be viewed as random effects. In other words, the external factors are accounted for by virtue of the fact that type-shift processes are construed as stochastic: roughly speaking, it is not assumed that a transition from a preferred type towards a dispreferred one is impossible, but only that the probability of such a transition is lower than the probability of the reverse transition. Such an assumption is necessarily implied by the notion of distributional universal in any event, although this implication is not always recognized, simply because type-shift processes constitute the only essentially linguistic process which affects the corresponding A-distribution and thus can, in principle, bring about a distribution which would be motivated by linguistic principles. If type-shift processes are not linguistically motivated, no typological distribution can ever reflect anything like linguistic preferences.

To sum up, the notion of distributional universal can only be defined as the stationary distribution of a type-shift process. This definition seems to capture the linguistic intuitions that underlie the notion of distributional universal but avoids the false assumptions inherent in the "languages-as-trials" approach.

2.3. *The ergodic hypothesis and interdependencies between typologies*

Statistical typological studies implicitly assume that the current language population (or, to be more precise, each member of this population) has already achieved the stationary distribution of linguistic properties (William Croft, personal communication). For instance, this assumption underlies the Relative Time Hypothesis put forward by Hawkins (1983: 256), which states, in effect, that the probability to find each specific language in a given type-state at a ran-

domly selected moment of its history is equal to the probability for a randomly selected member of the current population to belong to that type. Furthermore, it is only under this assumption that the only problem in establishing distributional universals is constituted by random effects of the birth-and-death process, as implied in all methodological discussions of the issue (cf. Section 4). On the other hand, Lass rejects the very possibility that a stationary distribution of linguistic properties can be achieved. In his model, the state of a language ALWAYS depends on the initial conditions, since languages are not ERGODIC systems, "at least in the temporal perspective that we have available" (Lass 1997: 302).

The ergodic property means that the system can return to any possible state and the expected time interval between two "visits" of a same state is finite. A stationary distribution exists only for ergodic systems. In somewhat anthropomorphic terms, the ergodic property can be thought of as the possibility for a language to "explore" all available typological options and return to any of them within a finite time interval. It is only under this condition that the notion of the probability for a language to "choose" one of these options makes sense. It may seem that the hypothesis of strong connection, which Lass does adopt (1997: 302), necessarily entails the ergodic property. This is indeed the case, but only under the assumption that the number of possible states is finite.[4]

At first sight, this condition is met by any typology. However, it should be taken into account that the notion of stationary distribution implies that a typology can be appropriately described as a MARKOV CHAIN, that is, that the transition probabilities are uniquely determined by the current type of a language (within the same typology). Most essentially, type-shifts within each typology are construed as events which are independent of any other linguistic properties. In principle, an interdependency between two or more typologies does not prevent the model from being applicable. It is just necessary to construct a more complex typology **S**, defined as a combination of all interdependent parameters, so that each type of **S** would constitute a subset of a certain type in each original typology **T**. The stationary probabilities for T_i can then be defined on the basis of the stationary distribution for **S** simply by the summation of the stationary probabilities of all types S_j that constitute subsets of T_i:

$$(2) \qquad p(T_i) = \sum_{S_j \subseteq T_i} p(S_j)$$

However, interdependencies between typological parameters are likely to increase the time period needed to achieve the stationary distribution along each parameter. Roughly speaking, the state of a system can be assumed to be independent of its initial state only after a time period that is sufficiently long for a language to visit all possible states, whereas each interdependency increases the total number of states to be visited. If a "typology" **S** constructed in an

attempt to account for all interdependencies were to include all parameters of language variation, then instead of a limited set of "type-states" we would be confronted with an infinite set of "all possible states" of a language. In this case, the ergodic hypothesis would be implausible: it hardly makes sense to assume that a language found in some state at some moment of time can return to precisely the same state within a finite time interval. This is the essence of Lass's claim that languages are not ergodic systems.

It is clear that the notion of distributional universals implies that some significant language properties can be adequately described in terms of a limited set of type-states, whereby the number of types remains reasonably small even if interdependencies are taken into account: a distribution pattern can only be discovered if the total number of types is much less than the number of languages under investigation. Within this frame, the ergodic property of a type-shift process is ensured by the requirement of strong connection, hence the model suggested in Section 2.2 is applicable. Another question is whether the stationary distribution of linguistic properties is already achieved by the language population: this should be a matter of empirical investigation, rather than of theoretical assumptions (see Section 4.2). An important advantage of the model advocated here is that it opens up the possibility of establishing distributional universals (in particular, to detect linguistically significant interdependencies) even if the currently existing languages do not display the stationary distribution of linguistic properties (Section 4.3).

3. The birth-and-death process and its impact on A-distributions

3.1. *Preliminary remarks*

As already mentioned, the birth-and-death process is commonly thought of as the major obstacle in establishing distributional universals. Since this process is essentially independent of the linguistic properties of the languages involved, it modifies A-distributions in a random way (Bell 1978: 171), yet the existence of LARGE LINGUISTIC FAMILIES suggests that these random effects may be statistically significant (Dryer 1989: 259–260). In spite of the disastrous implications of this hypothesis (cf. Section 1), there have been no attempts to estimate the potential impact of the birth-and-death process on A-distributions. This section is intended to show that the distorting effects of this process cannot have been as significant as is commonly believed, at least during the last several thousand years. An accurate assessment of the role played by this process suggests that some typological phenomena which have been interpreted as random effects of "historical accidents" may in fact reveal a drift of the language population towards stationary distribution.

The mathematical details of the research reported in this section will hardly be interesting for most readers, yet they are essential for those who might wish

to verify the results or to repeat the calculations for other values of relevant parameters. Therefore, the details are described in the Appendix; what follows here is only a brief overview of the facts needed to understand the results presented in the body of this section (the figures preceded by "A" refer to formulas provided in the Appendix).

Under the assumptions outlined in Section 2.1, the birth-and-death process in a language population can be modeled as a Feller-Arley process (A1). This model fully describes how the size of a population varies with time; more specifically, for an ancestor population of any given size, it is possible to calculate the distribution of population size $N(t)$ by the end of a time interval t, i.e., the range of possible values of $N(t)$ and their probabilities, see (A2)–(A4). The model also gives explicit formulas for the mean value (A5) and the variance (A6) of $N(t)$. Now the frequency of any linguistic type can be represented as a function of two independent variables each of which is described by the model, the number of languages which belong to this type, and the number of all other languages (A8). Hence, it is possible to examine how the frequency of a linguistic type in a population varies with time by virtue of the birth-and-death process in that population, that is, under the assumption that only this process is at work, cf. (A10)–(A12). More specifically, it is possible to figure out whether the frequency $f(t)$ of a linguistic type can significantly deviate from its frequency f_0 in the ancestor population by the end of a given time interval t.

3.2. *Parameters of the process*

A birth-and-death process in a population is determined by two parameters, which, for the purpose of this presentation, can be thought of as probabilities for a language to split and to die within a time interval of 100 years, p(B) and p(D) (see Section 3.1). In the model adopted here, the values of these parameters are constants characterizing the birth-and-death process in the entire population over a protracted period of time. This assumption requires two comments. First, it is obvious that both probabilities crucially depend on a variety of extra-linguistic factors (historical, geographical, etc.). However, inasmuch as these factors themselves are independent of the linguistic properties of a language, they can be neglected (of course, as far as the potential effects of the birth-and-death process on the A-distributions are concerned). Secondly, the assumption of constant p(B) and p(D) is evidently false if we wish to consider the entire history of the population, if only because at the present time the number of languages decreases, i.e., p(D) is greater than p(B). It is hardly necessary to argue that this could not have been the case for the entire history of humankind. The assumption of constant p(B) and p(D) is only plausible for time intervals which are relatively short; in the present pa-

per, the model is applied to time intervals of no more than several thousand years.

The best way to demonstrate the plausibility of this assumption would be to check the model against some actual data, for example, to compare variations in family size as predicted by the model and as observed in actual practice. The problem is that there are no empirical data which would be straightforwardly comparable with the predictions of the model. The model predicts that family size will be distributed exponentially, i.e., the probability that a genetic grouping of a certain time depth will have more than n members is given by the following formula:

$$(3) \qquad P_n = \left(1 - \tfrac{1}{m}\right)^n,$$

where m is the average size of genetic groupings of the SAME TIME DEPTH. On the other hand, the available classifications of languages can hardly be interpreted as representing the actual genetic structure of the population which could be accurately mapped onto the temporal scale. It is therefore to be expected that variations in the size of groupings at the same level of a classification do not follow the predicted exponential distribution.

This phenomenon is illustrated in Table 1, where data from three available classifications (Grimes (ed.) 1997, Voegelin & Voegelin 1977, Ruhlen 1987) are compared with the exponential distribution. The table provides figures for three values of n which are defined with respect to the mean value m: the probability for a family to be larger than the average ($n = m$), more than twice as large ($n = 2m$), and more than four times as large as the average ($n = 4m$), whereby the mean size of a major genetic grouping ranges from ca. 50 in *Ethnologue* (Grimes (ed.) 1997) to ca. 200 in Ruhlen (1987). The last row shows the corresponding figure for the exponential distribution. Whereas for $n = 2m$ the prediction lies within the range determined by the "actual" figures, it does not for the two other values: the model appears to underestimate both the number of small groupings (i.e., the predicted probability for a family to be larger than the average is higher than suggested by the classifications) and the number of very large groupings (with a size more than four times as large as the average). As already mentioned, these discrepancies are to be expected. Most importantly, in order to compare figures derived from existing classifications with the predictions of the model, it is necessary to make allowances with regard to the range of time depths, that is, to take into account that classifications comprise groupings of various time depths (as revealed by the simple observation that classifications characterized by the mean values of family size as different as 50 and 200, and thus obviously corresponding to different average time depths, contain a number of common families).

For the sake of simplicity, let us assume that the "major groupings" in each classification fall into two sub-sets corresponding to two different time depths.[5]

Table 1. *Frequencies of "large families" in major classifications of languages*

	Mean family size m	Frequencies of families with more than n members		
		$n = m$	$n = 2m$	$n = 4m$
Ethnologue (1997)	$m = 52$	0.15	0.09	0.05
Voegelin & Voegelin (1977)	$m = 102$	0.21	0.17	0.06
Ruhlen (1987)	$m = 201$	0.27	0.15	0.12
Exponential distribution		0.37	0.13	0.02

Table 2. *Predicted probabilities of "large families"*

Mean family size m	Assumed range of time depths	Predicted probability for a family to have more than n members		
		$n = m$	$n = 2m$	$n = 4m$
$m = 52$	5,000–7,500	0.28	0.14	0.05
$m = 102$	5,500–8,500	0.26	0.15	0.06
$m = 201$	7,000–9,500	0.28	0.14	0.05

The predictions of the model under this assumption are shown in Table 2. It can be easily observed that the correspondence between the predicted and the actual values is now much better: for $n = 4m$, the predicted figures are now precisely the same as those derived from *Ethnologue* and Voegelin & Voegelin (1977). The values for $n = m$ are still higher than in Table 1, which can be traced back to the existence of unclassified languages. Notably, since Ruhlen (1987) draws a distinction between "unclassified languages" (which are not taken into account here) and "language isolates", the predicted and the calculated figures for $n = m$ are virtually identical in this case. Thus, once we take into account the properties of existing classifications (even in an admittedly simplistic fashion), the model succeeds in making rather accurate predictions of the actual variation in size between genetic groupings. Table 2 also shows the assumed ranges of time depths for each classification, which, to the best of my knowledge, roughly correspond to the received temporal estimates.

Evidently, the temporal estimates given in Table 2 are based on some specific values of parameters p(B) and p(D) (cf. Note 5), which brings us to the next question, namely, how can these values be estimated? In order to determine the actual values of p(B) and p(D), it is necessary to know how many languages existed several thousand years ago. Yet the only figure that can be estimated is the number of ancestor languages that have at least one descendant in the current population (that is, the number of genetic groupings of a certain time

Table 3. *Parameters of the birth-and-death process in the language population*

HPS[1]	PBP[2]		Time depths of major genetic groupings			MSD[3]
	p(B)	p(D)	Ethnologue (1997)	Voegelin & Voegelin (1977)	Ruhlen (1987)	
450	0.079	0.009	4,500–6,500	5,000–7,500	6,500–8,500	0.03
600	0.097	0.035	5,000–7,500	5,500–8,500	7,000–9,500	0.03
1,000	0.136	0.087	6,000–9,500	7,000–11,000	9,000–12,500	0.03
2,000	0.208	0.178	10,000–15,500	11,500–18,000	15,000–20,000	0.03
4,000	0.307	0.296	27,000–42,000	30,000–50,000	40,000–55,000	0.03
5,500	0.362	0.360	$1.2{\cdot}10^6$–$1.9{\cdot}10^6$	$1.4{\cdot}10^6$–$2.3{\cdot}10^6$	$1.8{\cdot}10^6$–$2.5{\cdot}10^6$	0.03
5,900	0.375	0.375	$6.6{\cdot}10^6$–$1.0{\cdot}10^7$	$7.5{\cdot}10^6$–$1{,}2{\cdot}10^7$	$9.6{\cdot}10^6$–$1.3{\cdot}10^7$	0.03
5,950	0.377	0.377	$1.3{\cdot}10^7$–$2.0{\cdot}10^7$	$1.5{\cdot}10^7$–$2.4{\cdot}10^7$	$1.9{\cdot}10^7$–$2.6{\cdot}10^7$	0.03

[1] HPS: Hypothetical population size 3,700 years ago
[2] PBP: Parameters of the birth-and-death process
[3] MSD: Maximum standard deviation of a language frequency

depth); one can only guess how many members of the ancestor population have no surviving descendants. In what follows, I assume that the current population size is ca. 6,000, the number of genetic groupings of a time depth of ca. 3,700 years is estimated as 400 (which is the mean value of the estimates given by Bell (1978: 148) and Dryer (1989: 269), i.e., 478 and 322 respectively). Table 3 shows the corresponding values of p(B) and p(D) for various hypotheses on the population size 3,700 years ago, ranging from 450 to 5,950.[6] The next three columns of the table give the estimated range of time depths for major genetic groupings in *Ethnologue* (Grimes (ed.) 1997), Voegelin & Voegelin (1977), and Ruhlen (1987) for each pair of p(B) and p(D). The wild figures in the last lines should not be taken too seriously, since, as already mentioned, the model with constant p(B) and p(D) only works for time periods of several thousand years. They do indicate, however, that the last 4,000 years or so have seen a significant growth of the language population, that is, it is unlikely that the population size has been relatively stable during the last 4,000 years.

What is more important, however, is that the choice of p(B) and p(D) does not affect the major results, which in fact do not require any specific assumptions of this kind. That is to say, the estimates for deviations of linguistic frequencies induced by the birth-and-death process during the last several thousand years are roughly equal for all possible values of p(B) and p(D). This is demonstrated by the last column of Table 3, which shows that the predicted maximum of the standard deviation of frequency[7] remains roughly the same (ca. 0.03) independently of the adopted hypothesis. As will be clear from the following discussion, this somewhat unexpected phenomenon is accounted for

by a combination of two factors. Generally, the statistical effects of the birth-and-death process on the A-distribution depend on the ratio $p(\mathrm{B})/p(\mathrm{D})$: the less the value of $p(\mathrm{B})/p(\mathrm{D})$, the more significant the possible deviations of frequencies. On the other hand, the significance of such random effects rapidly decreases with the growth of the ancestor population. Yet our hypotheses on the size of the ancestor population and on the parameters of the process are not independent: the smaller the assumed size of the ancestor population, the higher the corresponding ratio $p(\mathrm{B})/p(\mathrm{D})$, so that the two relevant factors compensate for each other.

As a result, it turns out to be possible to suggest parameter-independent analytical estimates for deviations of frequencies induced by the birth-and-death process during the last several thousand years (cf. (4)–(6) below). In some cases, however, it will be necessary to use the results of computer calculations, which can only be based on some specific values of $p(\mathrm{B})$ and $p(\mathrm{D})$. The results presented in Sections 3.3 and 3.4 make use of two pairs of $p(\mathrm{B})$ and $p(\mathrm{D})$ which appear most plausible to me (lines 2 ($N_0 = 600$) and 3 ($N_0 = 1{,}000$) in Table 3).

3.3. *The dependency of birth-and-death effects on population size*

The potential impact of the birth-and-death process on A-distributions proves to depend most crucially on the size of the ancestor population: in small language populations birth-and-death effects can be highly significant, yet as the language population grows larger, these effects gradually become negligible.

As a first illustration of this dependency, let us consider the probability that the birth-and-death process will bring about a strongly uneven A-distribution in a descendant population. For the sake of simplicity, it can be assumed that an A-distribution for a two-way typology is "strongly uneven" if one of the types is more than twice as frequent as the other. Figure 1 shows the probability that such a distribution will be produced by the birth-and-death process in a small language population during a time period of ca. 4,000 years ($p(\mathrm{B}) = 0.097$, $p(\mathrm{D}) = 0.035$), under the assumption that the frequencies of the types in the ancestor population are equal. It can be easily observed that for an ancestor population of two languages (e.g., two descendants of the proto-language, one of which has acquired a new trait), this probability is close to 0.9, that is, it is much more probable that a descendant population will display a clear "preference" for one type than that the types will maintain similar frequencies. As the population grows larger, this probability rapidly decreases. If the initial number of languages is ten, it is already lower than 0.5 (but still fairly high). For a population of 70 members, the probability of such a strong bias in a descendant population falls below 0.05, for ancestor populations comprising more than 100 languages, this probability is lower than 0.01, that is, a strong

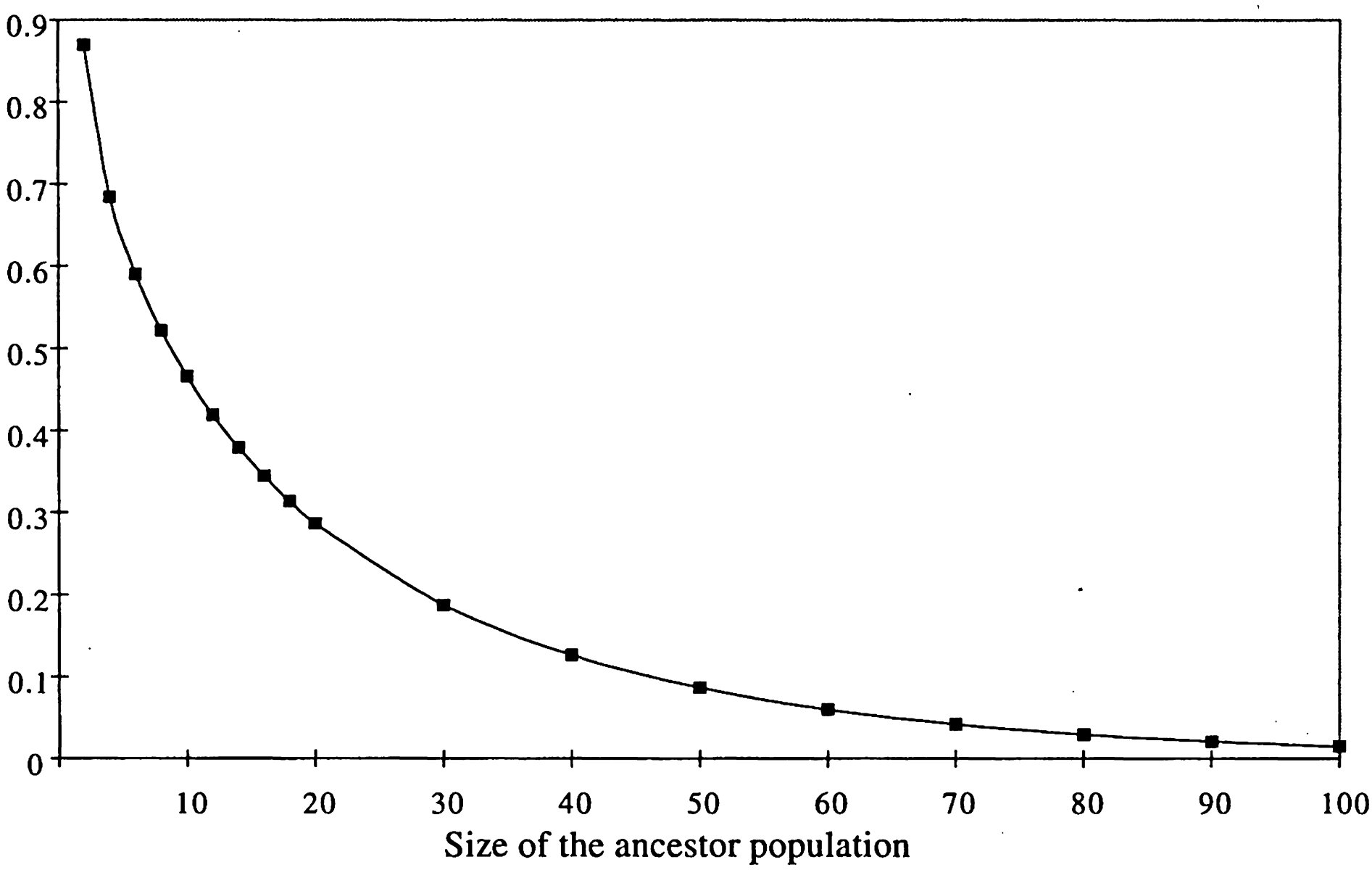

Figure 1. *Probability of a strong bias in a descendant population* ($p(\mathrm{B}) = 0.097$, $p(\mathrm{D}) = 0.035$, $t = 4000$)

bias induced by the birth-and-death process can be safely assumed to be impossible.

For larger populations, it will be more convenient to use a subtler parameter, namely, DEVIATION $d(t)$ of $f(t)$ from its initial value f_0. Since the birth-and-death process works independently of linguistic properties, the expectation of $f(t)$ is equal to f_0; cf. (A9). The standard deviation of $f(t)$ for large ancestor populations can be estimated as follows:

$$\sigma \approx \frac{1}{2\sqrt{K}} \tag{4}$$

where K is the number of ancestor languages that have at least one descendant in the current population; cf. (A12), (A7). Under the assumption of a normal distribution, which is quite plausible for large populations, the probability that an actual deviation $d(t)$ will exceed 2σ is less than 0.05 (with a level of confidence of 0.01, the actual deviation will be less than 3σ). The plausibility of such estimates is confirmed by a comparison with the calculated exact values, as illustrated by Figure 2 for two different pairs of parameters $p(\mathrm{B})$ and $p(\mathrm{D})$. The chart shows the calculated maximal value of a random deviation $d(t)$ after 4,000 years for a level of confidence of 0.05 and various sizes N of ancestor populations and estimates obtained on thè basis of (4) under the assumption of normal distribution. Note that the horizontal axis shows the number of ALL members of the ancestor population, including those that have no descendants

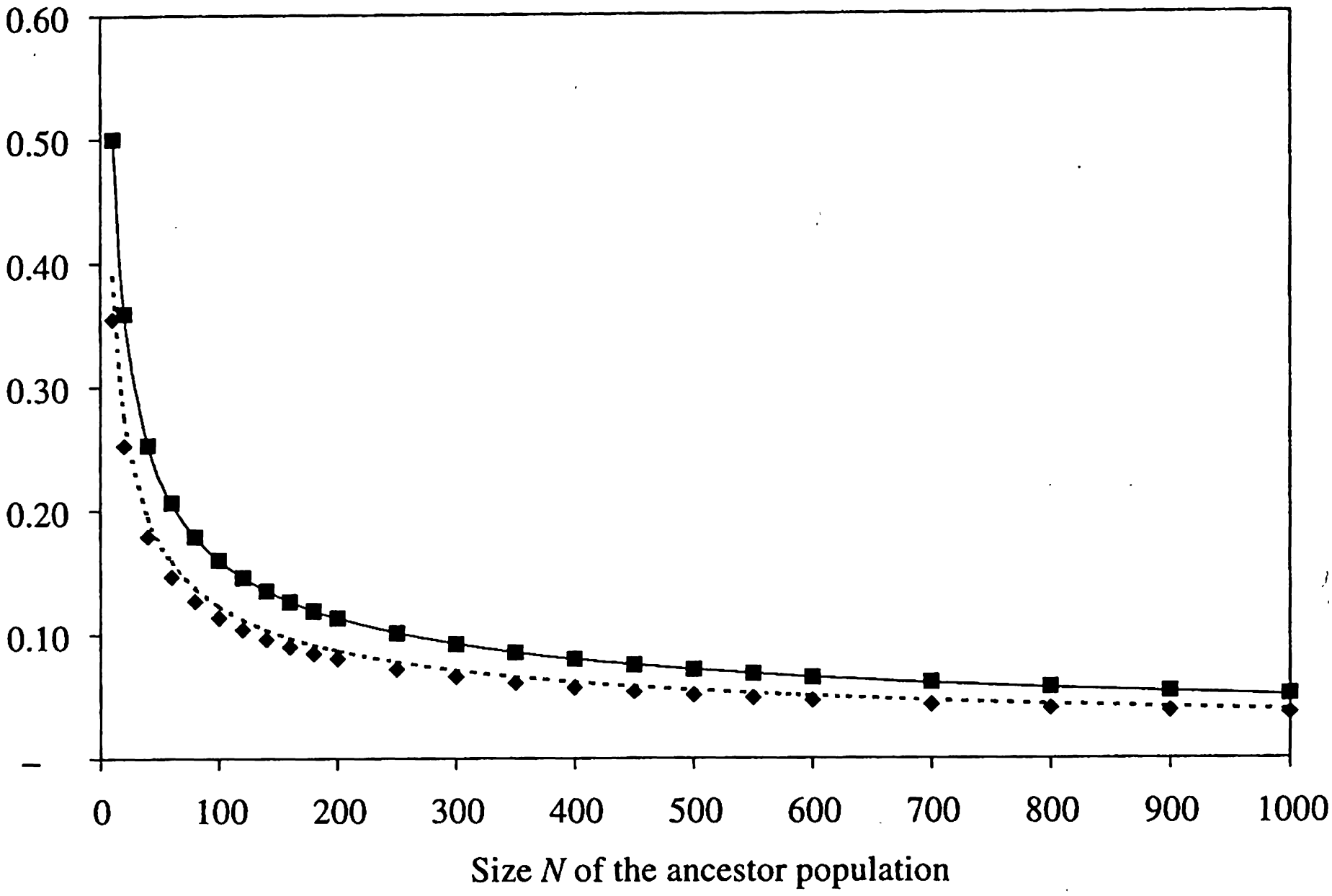

Figure 2. *Dependency of a random deviation d(t) on size N of the ancestor population (level of confidence* 0.05, $t = 4000$*)*
Legend:
black squares: exact values of $d(t)$ *for* $p(\mathrm{B}) = 0.136$, $p(\mathrm{D}) = 0.087$
black diamonds: exact values of $d(t)$ *for* $p(\mathrm{B}) = 0.097$, $p(\mathrm{D}) = 0.035$
solid line: estimated values $d(t) = 2\sigma$ *based on expression (4) for* $p(\mathrm{B}) = 0.136$, $p(\mathrm{D}) = 0.087$ *(*$K = 0.39 \cdot N$*)*
dashed line: estimated values $d(t) = 2\sigma$ *based on expression (4) for* $p(\mathrm{B}) = 0.097$, $p(\mathrm{D}) = 0.035$ *(*$K = 0.66 \cdot N$*)*

(the value of K was calculated according to (A7) for each pair of parameters). Two relevant facts are easy to observe: calculated values do not exceed the corresponding estimate and the maximum of actual deviation rapidly decreases with the growth of the population. Thus, the following rule of thumb can be used to estimate the significance of deviations induced by the birth-and-death process in large populations:

(5) With a probability of more than 0.95, the deviation $d(t)$ of frequency $f(t)$ in a descendant population from its frequency f_0 in the ancestor population is less than $\frac{1}{\sqrt{K}}$, where K is the estimated number of genetic groupings of time depth t.

For example, if K for the time depth of 3,500 to 4,000 years is estimated as 400, the deviations of frequencies induced by the birth-and-death process during this period of time can be assumed to be lower than 0.05.

3.4. *Actual deviations for small initial frequencies in large populations*

As shown in Section 3.3, the variation of A-distributions induced by the birth-and-death process in large language populations can be considered insignificant for relatively frequent language types. In fact, if an actual deviation of frequency does not exceed 0.05, it can be neglected if the frequency itself is high. Yet if the initial frequency is low, the same deviation may emerge as highly significant. In particular, if the initial frequency happened to be lower than 0.05, this estimate suggests that the type can easily disappear as a result of the birth-and-death process.

In reality, however, actual deviations $d(t)$ for small frequencies are much lower. This is because the rough estimate for the standard deviation proposed in (4) is close to its actual value only for initial frequencies close to 0.50. For low and high frequencies, the standard deviation is significantly less than suggested by (4), as shown by the following more precise estimate (6), which takes into account not only the population size, but also the initial frequency of a type; cf. (A12).

$$(6) \qquad \sigma \approx \sqrt{\frac{f_0(1-f_0)\alpha}{K}}, \text{ where } \alpha = 1 + \frac{p(\mathrm{D})}{p(\mathrm{B})}$$

It can be easily observed that σ achieves the maximum for $f_0 = 0.50$. The greater the difference between f_0 and 0.50, the smaller the standard deviation. The additional coefficient reveals the dependency of deviations on the ratio $p(\mathrm{D})/p(\mathrm{B})$ (cf. Section 3.2). The dependency of $d(t)$ on the value of f_0 is shown in Figure 3 for $K = 400$ and two different pairs of parameters $p(\mathrm{B})$ and $p(\mathrm{D})$ (the level of confidence is 0.05). The figure also shows estimate $d(t) \leq 2\sigma$ obtained on the basis of (6) for $\alpha = 1.5$ (that is, under the assumption that $p(\mathrm{B})$ is at least twice as large as $p(\mathrm{D})$, cf. Table 3).

Although deviations of frequency induced by the birth-and-death process in large populations are always small with respect to the initial value of frequency, this does not eliminate the problem of very rare types: there exists some probability that such a type will cease to exist due to the birth-and-death process, which may seem to constitute a linguistically very significant modification of an A-distribution. This probability proves to be extremely small for large populations: for all possible values of $p(\mathrm{B})$ and $p(\mathrm{D})$, the probability that a language type has died out within the last 4,000 years exceeds 0.005 (sic!) only if its frequency in the ancestor population was less than 0.006, that is, if it was represented by less than 1 % of languages (these figures are based on expression (A3); the values of parameters are taken from Table 3). What is more important, however, is that even such cases cannot be considered as really significant modifications of A-distributions. This potentially controversial claim is but another formulation of a well-known point: any empirical absolute universal stating that a certain language type is impossible is in effect a

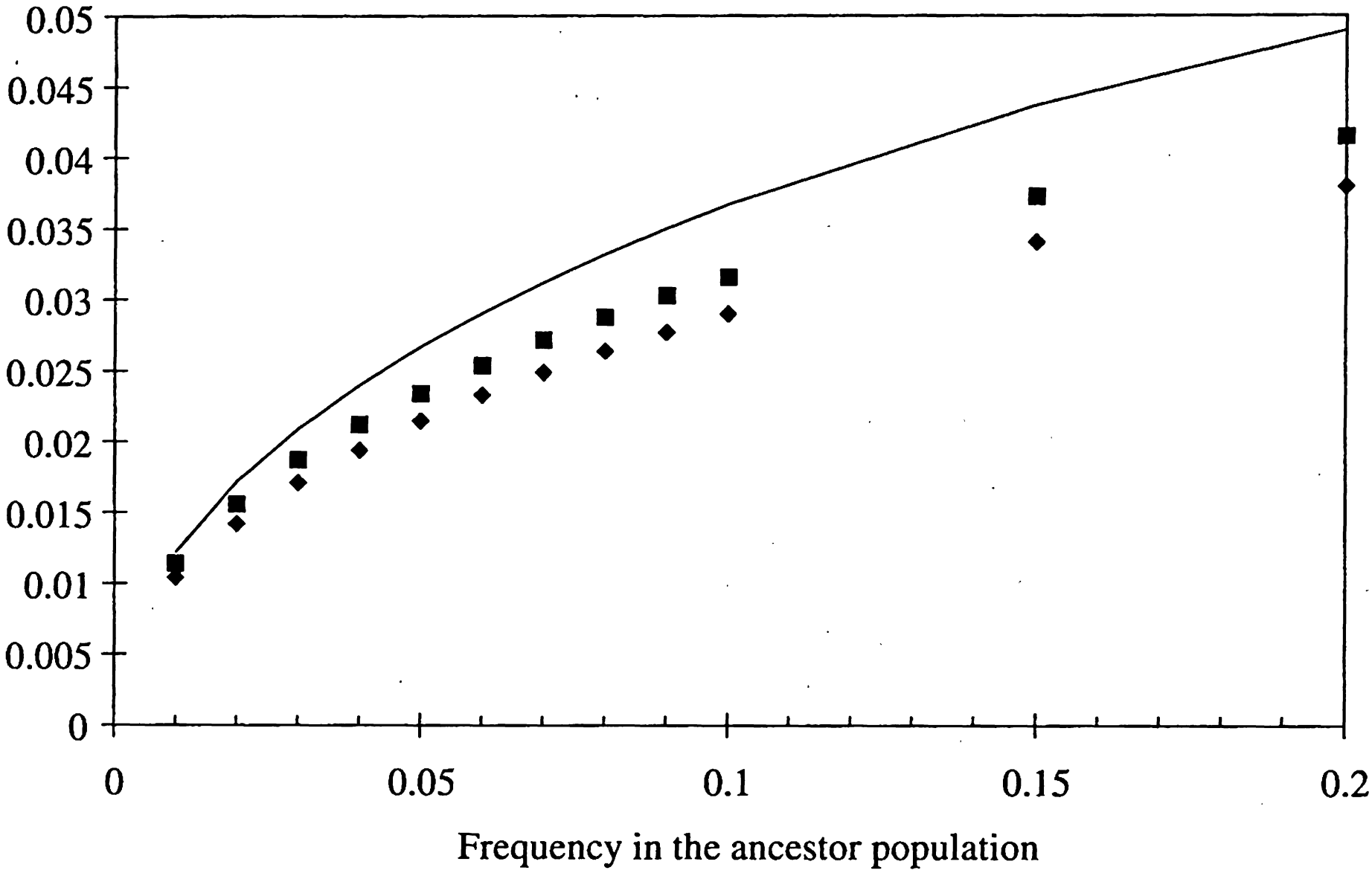

Figure 3. *Dependency of random deviation d(t) on the initial value of frequency (level of confidence* $0.05, K = 400, t = 4000$*)*
Legend:
black squares: exact values of $d(t)$ *for* $p(\mathrm{B}) = 0.136, p(\mathrm{D}) = 0.087$
black diamonds: exact values of $d(t)$ *for* $p(\mathrm{B}) = 0.097,\ p(\mathrm{D}) = 0.035$
solid line: estimated values of $d(t) = 2\sigma$ *based on expression (6) for* $\alpha = 1.5$

distributional universal stating that this type is very rare. In fact, even if some type with an initial frequency below 0.006 were to retain precisely the same frequency in the current population, it would hardly be detected by a sampling procedure. Thus, if some logically possible type is not attested in a large set of languages, it can only mean that its probability is very low, but not that this probability is zero; that is, a zero frequency and a very low frequency must entail virtually the same linguistic conclusions.

3.5. *Interpretation*

The estimates presented in this section can be summarized as follows: in large populations, the birth-and-death process can be neglected, which means that all statistically significant modifications of A-distributions are to be attributed to type-shift processes in the history of specific languages. On the other hand, in small populations, birth-and-death effects are so dramatic that type-shift processes can hardly play a significant role.

Now it seems reasonable to assume that the actual history of the language population has known both types of periods. On the one hand, it is clear that

the population has been large during the last several thousand years, i.e., within the time period when the currently observable genetic groupings came into existence. Hence, the initial population (as defined in Section 2.1 on the basis of "evolutionary" considerations) was certainly large enough to rule out any further statistically significant random effects. On the other hand, at the earliest stages of its history, the language population was probably small, or even comprised only a single proto-language. The time interval lying between these stages is to be counted in thousands of centuries, and we do not know when exactly the population crossed the borderline between being "small" and "large" (this may even have happened more than once). As far as the initial population is concerned, the only conclusion possible under these circumstances is that there can be no rational assumptions with regard to its linguistic properties, if only because it might have been strongly biased by the "start-up" effects of the birth-and-death process, let alone unknown properties of the proto-language(s).

Thus, the model proposed in Section 2.1 can now be revised: the birth-and-death process can have biased INITIAL A-distributions in a number of linguistically unmotivated and highly significant ways, whereas any further significant modifications of A-distributions should be interpreted as resulting mainly from type-shift processes. For instance, the frequency of SVO languages is now ca. 41.79 % (Tomlin 1986: 22), whereas Dryer (1989: 269–270) estimates this frequency for the time depth of genera (i.e., 3,500 to 4,000 years ago) as ca. 26 % (57 genera in a sample of 218) and interprets this difference as a result of historical accidents. This interpretation is in an obvious contradiction with the estimates obtained in this section: assuming the number of genera to be no less than 322 (which is Dryer's own estimate), the actual deviation induced by the birth-and-death process during the relevant time period exceeds 0.056 with a probability of less than 0.05. It follows that the attested increase of frequency (ca. 15 %) should be attributed first and foremost to type-shift processes.

This means that the linguistic interpretation of such phenomena should be precisely opposite to that suggested by Dryer. He takes the A-distribution as reconstructed for the level of genera as a more appropriate estimate of the hypothesized distributional universal than the current A-distribution. Yet since the difference between these distributions results mainly from transitions between types, it is likely to reveal a drift of the population from the initial A-distribution towards the stationary distribution (see Section 2.2). If so, then the current A-distribution is likely to be closer to the distributional universal than any of its earlier counterparts. This observation shows that an unjustified overestimation of random birth-and-death effects can be no less harmful for linguistic conclusions than their underestimation.

4. How to establish distributional universals?

4.1. *The problem of non-independence: a new perspective*

The estimates presented in the previous section may appear to remove the major obstacle in establishing distributional universals, i.e., the "distorting effects of large families" (Dryer 1989: 260). In reality, however, these results show that the problem resides elsewhere: the current A-distributions need not be independent of their initial counterparts. In particular, they may still bear statistically significant traces of those birth-and-death events that had happened before t_0, i.e., when the language population had been small.

The problem of random "start-up" effects and their long-term consequences has been discussed in the literature, yet only in terms of idiosyncratic features of proto-languages which might have been inherited by their currently existing descendants (Maddieson 1991: 352). The birth-and-death process emerges as an additional source of such effects: even if the language population originated from a few independent proto-languages, the problem is not eliminated, at least insofar as the proto-population is assumed to have been relatively small. On the other hand, the long-term consequences of any "start-up" phenomena cannot be reduced to instances of the RETENTION of linguistic properties – contrary to what is commonly implied in the literature (cf., e.g., Vennemann 1992: 48). As discussed in Section 2.3, the current type of a language can be assumed to be statistically independent of its initial type only after a time interval which is sufficiently long for a language to be able to visit all possible type-states, not just to change from one state towards another.

The ultimate solution of the problem of distributional universals would depend on whether the time period separating the current population from its initial counterpart has been sufficiently long for type-shift processes to bring about the stationary distribution. If this were the case, the solution would be quite straightforward: the current A-distributions would provide most accurate estimates of distributional universals. In other words, an approach like that adopted in Tomlin (1986) would be fully justified: the current A-distributions might be taken to reflect universal linguistic principles, although the random deviations induced by the birth-and-death process should be taken into account in estimating margin errors. Otherwise, no synchronic statistical evidence would suffice to establish a distributional universal (cf. Section 4.3).

4.2. *Tests for stationary distributions*

Recent typological studies have accumulated overwhelming evidence against the general assumption of the stationary distribution of linguistic properties in the current language population, although this evidence has hardly been interpreted in such terms.

To begin with, there is a relatively simple way to test whether a time interval *t* is sufficiently long to ensure that the current type of a language is independent of its initial type. If this were the case, then the INTERNAL A-distributions would be roughly similar for all genetic groupings with a time depth of no less than *t*. Indeed, if the current type of a language is statistically independent of its initial type, then, by definition of independence, the probability for a language to belong to this type must be the same for all possible initial types. The frequency of this type among the descendants of any ancestor language must reflect this probability, hence, it must be roughly the same for all genetic groupings. In effect, this means that the linguistic type of a language should be independent of its genetic affiliation. Yet this is generally not the case even for the major genetic groupings and relatively unstable linguistic phenomena (Perkins 1989: 305–311), that is, roughly speaking, for the estimated time depth of the initial population (cf. Section 2.1). Accordingly, the current population is bound to retain statistically significant traces of the initial A-distributions, at least for some typologies.

It may be the case, however, that the estimate of 10,000 years as the time depth of the initial population (cf. Section 2.1) is too pessimistic. That is to say, the language population might have achieved a sufficient size and the current level of evolution earlier than 10,000 years ago; if so, then the time period available to achieve the stationary distribution has been longer than the time depth of major genetic groupings. In this case, the dependency of linguistic properties of a language on its genetic affiliation is not enough to reject the assumption of a stationary distribution. Another argument against this assumption, which is not sensitive to the estimated time depth of the initial population, is given by the phenomenon of large linguistic areas (Dryer 1989), i.e., roughly speaking, by the fact that the language populations of different continents generally exhibit significantly different A-distributions. Yet if the stationary distribution had been achieved, then the probability to belong to a certain type would be identical for all languages, hence, the A-distributions in all sub-populations would be roughly similar. In other words, the phenomenon of large linguistic areas indicates that the A-distributions in each large sub-population still depend on the distribution of linguistic properties in its own ancestor population, which means that the stationary distribution has not been achieved. Since the somewhat mysterious phenomenon of large areas attracts more and more attention in the literature, it is worth noting that it receives a straightforward account within the model suggested here.

The attested dependencies of linguistic properties of a language on its genetic and areal affiliation show that the general assumption of a stationary distribution does not hold, which means that this assumption must be tested for each specific typology. For example, if the internal A-distributions for some typology are demonstrably similar for a representative set of genetic group-

ings, then these A-distributions can be employed to estimate the stationary distribution and hence the distributional universal. The same is true if the A-distributions are similar for a representative set of (areal) sub-populations. The latter test looks similar to that suggested by Dryer (1989). However, there are two major differences. First, the test proposed here is based on CURRENT area-internal distributions, whereas Dryer considers distributions at the level of genera (cf. Section 3.5). Secondly, it is required that the A-distributions under comparison reflect the same PROBABILITIES (e.g,. according to the χ^2-test), not just that the frequency of one type be consistently larger than the frequency of another (as in Dryer's test).

It might be argued that these tests are not likely to give positive results for any typology, given that the aforementioned dependencies have been attested for relatively unstable word order phenomena. For more stable linguistic phenomena, the time period needed to achieve the stationary distribution is likely to be even longer. This is not necessarily the case, since various word order parameters are known to be mutually dependent, which means that the type-shift process for word order typology (cf. Section 2.3) comprises a relatively large number of states. Since the time period needed to achieve the stationary distribution increases with the number of possible type-states, this period can be shorter for smaller independent typologies (if any). To sum up, although the language population cannot be assumed to have achieved the stationary distribution of all linguistic properties, this assumption may still prove plausible for some of them.

4.3. *Estimating the transition probabilities*

If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data. It is impossible to escape this problem by means of one or another sampling technique, as suggested, e.g., in Perkins (1989): even if a sample were to include a single descendant of each member of the initial population, the distribution of linguistic properties in such a sample would nonetheless be determined to some extent by the initial A-distribution, since the linguistic properties of EACH language are not independent of the linguistic properties of its ancestor. In this case, the only way to discover a distributional universal is to estimate transition probabilities and as it were to "predict" the stationary distribution on the basis of the equations in (1).

The idea of looking at something like transition probabilities is by no means new for research on language universals. Suffice it to recall that Alan Bell suggested in 1978 (with a reference to a personal communication with Joseph Greenberg) that "a case can be made that such research [on language universals; EM] can properly be conceived as sampling language changes, not lan-

guages themselves" (1978: 146). However, the sampling method proposed by Bell cannot, in principle, provide estimates of transition probabilities, since it gives information only on the current state of languages (Bell 1978: 147–148). It is clear, however, that if we want to estimate the probability of a shift $T_i \rightarrow T_j$ we must compare two quantities: the number of languages which have undergone this shift and the number of languages which have retained type T_i or shifted to another type within the same time interval. Accordingly, in order to estimate a transition probability $p(T_i \rightarrow T_j)$ for some time interval t, one would need a sample of languages which can be assumed to have been in state T_i t years ago. The current frequency of type T_j in this sample would give an estimate of the transition probability $p(T_i \rightarrow T_j)$.

One way to obtain such estimates is to analyze internal A-distributions in genetic groupings of a relatively small time depth t. As mentioned in Section 2.2, t should be selected in such a way that a transition be possible but not highly probable. Thus, the majority of languages in each grouping would have the inherited type, but some would have changed towards another type. In this way, it is possible to construct a sample of pairs $<T_i; T_j>$, where T_i is the source type and T_j the target type, hence, to estimate the transition probability for each pair. Another possibility might be based on sampling shifts in progress; such cases emerge in virtually all typological studies and are commonly dealt with in terms of "intermediate" or "mixed" types. If a random language sample contains n cases of a shift $T_i \rightarrow T_j$ in progress then the transition probability $p(T_i \rightarrow T_j)$ can be estimated as $\frac{n}{N}$, where N is the total number of all clear instances of T_i and all instances of shifts from this type to any other type.

Within the framework introduced in Sections 2.2–2.3, a statistical analysis of a type-shift process can have three different results. First, a typology may prove not to be strongly connected. In this case, it has no associated distributional universal. Secondly, some or all transition probabilities may prove to vary depending on some other linguistic parameters; in other words, this procedure may lead to a discovery of an interdependency between different typologies. In this case, a distributional universal can only be established on the basis of a more complex typology, as described in Section 2.3. It is worth noting, however, that the discovery of such an interdependency is an interesting result in itself; furthermore, an interdependency established on the basis of transition probabilities cannot be induced by accidental properties of initial A-distributions, i.e., it is bound to be linguistically significant. Finally, if a typology proves to be strongly connected and independent, the distributional universal can be calculated on the basis of transition probabilities according to the equations in (1).

To conclude, the brief overview of possible empirical procedures presented in this section is by no means intended to underestimate difficulties involved in the statistical analysis of type-shift processes. My major goal has been to show

that a workable procedure of estimating transition probabilities is a necessary prerequisite for any claims of distributional universals for typologies which cannot be assumed to have achieved a stationary distribution, rather than to put forward a detailed proposal, not to mention specific typological applications. I believe that an approach to the statistical analysis of typological data cannot be verified or falsified by specific applications; it must be shown to be theoretically justified BEFORE it can be applied, and this is what I have attempted to do in the present paper. It is quite easy to anticipate that some of the assumptions involved in the approach advocated here will appear too strong once they are explicated, which is of course a sufficient reason to reject the proposed solution. It seems therefore worth stressing once more that the assumptions adopted here are substantially weaker than those implied in current typological practice (as well as in previous methodological proposals). It may well be the case that a more sophisticated stochastic model is required, but certainly not that the currently adopted practices are reliable. What I have proposed is essentially to do a single step from the evidently inappropriate model of independent trials towards a more appropriate and powerful model of Markov chains, which has proved to be adequate and quite fruitful for a variety of rather complex phenomena. It remains to be seen whether linguistic typology can also benefit from this step.

Appendix

The Feller-Arley process, or linear birth-and-death process, is a continuous-time Markov process with the following transition rates (Feller 1971: 454–457, Srinivasan & Mehata 1978):

$$\text{(A1)} \qquad \begin{aligned} q_{n,n+1} &= n\lambda \quad \text{for } n > 0 \\ q_{n,n-1} &= n\mu \quad \text{for } n > 0 \\ q_{n,m} &= 0 \quad \text{for } n \geq 0, m \geq 0, m \neq n \pm 1, \end{aligned}$$

where λ and μ are probability densities for birth and death respectively. Probabilities $p_n(t|1)$ for a language to have n descendants by the end of time interval t are given by the following expressions:

$$\text{(A2)} \qquad p_n(t|1) = (1-a)(1-b)b^{n-1} \quad \text{for } n > 0,$$

$$\text{(A3)} \qquad p_0(t|1) = a,$$

where $a = (\mu e^{(\lambda-\mu)t} - \mu)/(\lambda e^{(\lambda-\mu)t} - \mu)$, $b = \lambda a/\mu$. Probabilities for a population with initial size N_0 to have n members by the end of time interval t are:

$$\text{(A4)} \qquad p_n(t|N_0) = a^{N_0} b^n \sum_{j=0}^{\min(N_0,n)} \binom{N_0+n-j-1}{n-j} \binom{N_0}{j} \left(\frac{1-a-b}{ab}\right)^j.$$

The mean value of population size $N(t|N_0)$ and the variance are given by the following equations:

$$\text{(A5)} \qquad \mathrm{Exp}[N(t|N_0)] = N_0 e^{(\lambda-\mu)t},$$

$$\text{(A6)} \qquad \mathrm{Var}[N(t|N_0)] = N_0 \frac{\lambda+\mu}{\lambda-\mu} e^{(\lambda-\mu)t}(e^{(\lambda-\mu)t}-1).$$

The expected number of ancestor languages that will have at least one descendant is:

$$\text{(A7)} \qquad K(t|N_0) = N_0(1-p_0(t|1)).$$

The general expression in (A4) cannot be employed in computer calculations for large N_0, since the alternating series leads to a dramatic loss of precision. Therefore, the calculations used in the paper are based on expression (A2) for $N_0 = 1$, which must be convoluted numerically with itself N_0 times in order to obtain the value of $p_n(t|N_0)$.

The frequency $f(t|N_0, f_0)$ of a linguistic trait with initial frequency f_0 can be represented as a function of two independent variables distributed according to (A4), cf.

$$\text{(A8)} \qquad f(t|N_0, f_0) = \frac{N(t|N_0^+)}{N(t|N_0^+) + N(t|N_0^-)},$$

where $N_0^+ = N_0 f_0, N_0^- = N_0(1-f_0)$, under the condition that at least one language survives by the end of time interval t. The conditional distribution function for $f(t|N_0, f_0)$ is given by the following equation:

$$\text{(A9)} \qquad F(f;t,N_0,f_0) = P_S^{-1} \sum_{i\geq 0, j\geq 0, i+j>0} p_i(t|N_0^+) p_j(t|N_0^-) \vartheta(f - \frac{i}{i+j}),$$

where $\vartheta(x)$ is the step function, $P_S = 1 - p_0(t|N_0^+)p_0(t|N_0^-)$ is the probability that at least one language exists by the end of t.

According to the Central Limit Theorem, $p_n(t|N_0)$ for $N_0 \gg 1$ is a normal distribution with mean value (A5) and variance (A6). The expectation of frequency $f(t|N_0, f_0)$ is simply

$$\text{(A10)} \qquad \mathrm{Exp}[f(t|N_0, f_0)] = \frac{N_0 f_0}{N_0 f_0 + N_0(1-f_0)} = f_0.$$

The variance can be estimated as:

$$\text{(A11)} \qquad \mathrm{Var}[f(t|N_0, f_0)] \cong (\frac{\partial f}{\partial N_0^+})^2 \mathrm{Var}[N(t|N_0^+)] + (\frac{\partial f}{\partial N_0^-}) \mathrm{Var}[N(t|N_0^-)].$$

By substituting (A6) in (A11), we obtain the following estimate:

$$\text{(A12)} \quad \mathrm{Var}[f(t|N_0, f_0)] \cong \frac{\lambda+\mu}{\lambda-\mu} \frac{f_0(1-f_0)}{N_0} \frac{e^{(\lambda-\mu)t}-1}{e^{(\lambda-\mu)t}}$$

Estimate (6) for the standard deviation is derived from (A12) and (A7). The simplified estimate in (4) is the limit of $\mathrm{Var}[f(t|N_0, 0.5)]$ for $N_0 \to K$.

Received: 12 January 2000
Revised: 3 October 2000

Universität Bielefeld

Notes

Correspondence address: 5 Murray, # 311, San Francisco, CA 94112, USA.; e-mail: maslova@jps.net

The computer program used in the present paper has been written by Eugene Levin. His contribution to this research project can hardly be overestimated; our discussions of the parallels between the evolution of language populations and other random processes have, to a large extent, inspired the approach presented in this paper. I am grateful to Bill Croft, Matthew Dryer, Leonid Kulikov, Eugene Levin, Boris Maslov, Edith Moravcsik, Johanna Nichols, and Yakov Testelec for their comments on earlier drafts of this paper and to Bernard Comrie, Bill Croft, Östen Dahl, Andrej Kibrik, Christian Lehmann, Sergey Say, and, more generally, the audience of the Winter Typological School in Moscow and of my *Habilitationskolloquium* in Bielefeld for encouraging discussion and insightful questions. Special thanks are due to Eugene Levin, Vladimir Nekrutkin, and Igor Zolotukhin for their help in the analysis of the relevant mathematical results.

1. To give a simple example, I believe that no linguist would find it difficult to explain why verb-initial basic orders are preferred over subject-initial orders (of course, if such a distribution were observed). Hence, no explanation of why the actual distribution shows a higher frequency of subject-initial orders can prove that this distribution is non-accidental.
2. A time interval is "small" if a type-shift is possible but not highly probable (so that two subsequent shifts at one "step" of the process are virtually impossible).
3. The equations in (1) for the stationary distribution can be found in any introduction to probability theory (cf., e.g., Feller 1971: 392–394).
4. A system with an infinite number of states can also be ergodic, but this is a more complicated issue, which is outside the scope of the present paper.
5. More precisely, it has been assumed that each set of families comprises two subsets with different mean values m_1 and m_2 (60 % and 40 % of families respectively) and that the family size is distributed exponentially within each subset. The values of m_1 and m_2 have been selected in such a way as to match the mean family size for each classification ($m_1 = 20$, $m_2 = 100$ for $m = 52$; $m_1 = 30$, $m_2 = 210$ for $m = 102$; $m_1 = 80$, $m_2 = 382.5$ for $m = 201$). The corresponding ranges of time depths shown in Table 2 have been calculated under the assumption that $p(\mathrm{B}) = 0.097$, $p(\mathrm{D}) = 0.035$.

6. For any specific hypothesis on the population size 3,700 years ago, all other values presented in Table 3 can be calculated on the basis of (A4)–(A7). The temporal estimates are based on figures given in Note 5.
7. The standard deviation achieves the maximum for $f_0 = 0.50$, cf. (A11).

References

Bell, Alan (1978). Language sampling. In Greenberg, Ferguson, & Moravcsik (eds.) 1978, 125–156.

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

Croft, William (1995). Modern syntactic typology. In Shibatani & Bynon (eds.) 1995, 85–142.

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

Feller, William (1971). *An Introduction to Probability Theory and its Applications. Volume 1.* Corrected re-impression of 3rd edition. New York: Wiley.

Greenberg, Joseph H. (1978). Diachrony, synchrony and language universals. In Greenberg, Ferguson, & Moravcsik (eds.) 1978, 61–91.

— (1995). The diachronic typological approach to language. In Shibatani & Bynon (eds.) 1995, 143–166.

Greenberg, Joseph H., Charles A. Ferguson, & Edith A. Moravcsik (eds.) (1978). *Universals of Human Languages, Volume 1: Method & Theory.* Stanford: Stanford University Press.

Grimes, Barbara F. (ed.) (1997). *Ethnologue: Languages of the World (plus Supplement: Ethnologue Index).* 13th edition. Dallas: Summer Institute of Linguistics.

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

— (1990). Seeking motives for change in typological variation. In William Croft, Keith Denning, & Suzanne Kemmer (eds.), *Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on his 75th Birthday*, 95–128. Amsterdam: Benjamins.

Lass, Roger (1997). *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.

Maddieson, Ian (1991). Investigating linguistic universals. In *Proceedings of the XIIth International Congress of Phonetic Sciences, Volume 1*, 346–354. Aix-en-Provence: Université de Provence.

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

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

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

Ruhlen, Merritt (1987). *A Guide to the World's Languages, Volume 1: Classification*. London: Arnold.

Shibatani, Masayoshi & Theodora Bynon (eds.) (1995). *Approaches to Language Typology*. Oxford: Clarendon.

Srinivasan, S. K. & K. M. Mehata (1978). *Probability and Random Processes*. New Delhi: Tata McGraw-Hill.

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

Vennemann, Theo (1992). Language universals: Endowment or inheritance? *Diachronica* 9: 47–60.

Voegelin, Charles F. & Florence M. Voegelin (1977). *Classification and Index of the World's Languages*. New York: Elsevier.