Neolithic Linguistics

Søren Wichmann
Max Planck Institute for Evolutionary Anthropology & University of Copenhagen

0. Introduction[1]

The term 'neolithic linguistics' is not intended to connote some kind of particularly primitive approach to linguistics, but rather refers to my attempt in this paper to discern linguistic signatures in the present-day global language distribution of the innovation and spread of farming. More specifically, my aim is to seek a more precise way to formulate the general idea that there is a correlation between the distribution of the major language families in the world and areas of early, prehistoric cultivation. While Indo-European figures prominently in the literature as a language family possibly conforming to that hypothesis, it will simply be treated on a par with the other language families in the world in this paper. That is, the hypothesis is tested on a global scale. If it can be shown to work well in general, it should be taken seriously as a candidate for explaining the expansion of Indo-European. While I hope that my ideas will not strike the reader as primitive they do have a measure of simplicity. Many factors that may contribute to the growth and spread of language families, such as other technological and cultural advantages, success in warfare, or simple Wanderlust, are disregarded, not because they always are irrelevant, but because they seem to play minor roles in the over-all picture.

The hypothesis that there is a correlation between the spread of major language families on the globe and areas where early developments of farming have taken place has some early predecessors (Heine-Geldern 1932 and Romney 1957, cited in Bellwood and Renfrew 2003:xiv),

---

[1] Preliminary versions of this paper were presented at the conference "Indoeuropæerne - sproget og forhistorien", University of Copenhagen, March 7, 2002, as a guest lecture at the Dept. of Anthropology, Northern Illinois University, April 2, 2002, at the conference ARCLINGII, Canberra, Oct. 2, 2002, as an Institute Seminar lecture, Max Planck Institute for Evolutionary Anthropology, Nov. 7, 2003, and at the Workshop on Prehistoric Chronology: Language, Genes and Migrations, Santa Fe Institute, March 5, 2004. I owe thanks to many people in the various audiences for their comments, in particular Michael Lachmann, Roy King, Peter Peregrine, R.G. Matson, Svante Pääbo, Peter A. Underhill, Henry Wright, Sergei Starostin, and Cecil H. Brown.

but is mainly associated with relatively recent work by the archaeologists Peter Bellwood and Colin Renfrew (Bellwood 1994, 1997, 2001, 2003a; Renfrew 1987, 1994, 2000, 2003). Recently it was the focus of a major conference organized by the two just mentioned scholars, cf. Bellwood and Renfrew (2003). The idea that different neolithic transitions have driven the spread of major language families is, in part, inspired by Ammerman and Cavalli-Sforza's (1984) model of a demic diffusion, i.e. gradual demographic increase and population spread correlating with the advance of farming from a Middle East center of innovation up through Europe. According to Renfrew (1987), this demic diffusion can be correlated with the spread of Indo-European speakers. It has also stirred controversy. Some archaeologists maintain that for the periphery of the area, such as southern Scandinavia or the Circum-Balitic region, the spread of agriculture is better explained as a result of cultural rather than demic diffusio (Zvelebil and Zvelebil 1988, Price 1996). On the other hand, the application by Bellwood (1994, as well as later publications) of the scenario to the Pacific area, where there is a good correlation between the so-called Lapita culture and the great Austronesian language family, now seems to represent the received view, and so may count as a success story in the development of the language/farming dispersal hypothesis. From these two cases the idea has been generalized, and it is now claimed by Bellwood and Renfrew that it is true of many major language families in the world that their spread has been driven by prehistoric farming.

As an example of how the generalization of the idea has inspired people to look at things in a new way, the Uto-Aztecan language family of northern Mexico and the southwestern United States could be mentioned. Bellwood (1994, 2001) has claimed that the spread of this language family might also be seen at the product of demic diffusion driven by early farming. This, however, would require that the Uto-Aztecans originated in the south, i.e. in Mesoamerica where we find agriculture from very early on, and not from somewhere in the southwestern United States as traditionally supposed (Fowler 1983). Furthermore, instead of the scenario whereby some Uto-Aztecan groups have taken up farming in relatively recent times, we would have to accept a scenario according to which some groups have *given up* farming. Jane Hill (2001, 2003) has tried to support this new scenario by claiming proto-Uto-Aztecan ancestry for some words in the Hopi language that refer to the maize complex. The claim here is that these words have retained their original meanings in Hopi, whereas they have changed in other Uto-Aztecan languages, where they no longer refer to items related to the maize complex.[2] It is debatable

---

2 One might argue that it is futile to seek for a correlation between prehistoric farming and a given language family unless the inventory of proto-vocabulary for that family contains evidence for farming. This type of argumentation is put forward in recent work by Roger Blench, whose position is that "if you assert that the Niger-Congo phylum spread following the adoption of agriculture, then vocabulary in the actual languages of the phylum must support this assertion,

whether Hill's linguistic arguments are correct (Cambell 2003). On the other hand, there exists the possibility that Uto-Aztecan is genetically related to Mixe-Zoquean, a language family of Mesoamerica (Wichmann 1999, 2003). This would seem to support an out-migration from Mesoamerica. This is just one example of the kinds of issues that may arise in specific cases when one applies the farming/language dispersal hypothesis. My assessment is that even if the hypothesis raises questions in individual cases and may have possible exceptions, it is fruitful hypothesis, one which deserves to be examined closely and tested on a world-wide scale.

1. Problems with the language/farming hypothesis as originally formulated

An immediate problem that strikes one when examing the language/farming hypothesis is that in its current formulation it is rather vague. In the following programmatical statement by Bellwood (2001: 182) it is said that "certain major language families" exhibit a correlation with prehistoric farming. Bellwood does not give criteria that would allow one to single out *the* major language families in question from other major language families.

> Human prehistory gives us a record of two very important, yet at first sight unrelated, examples of expansion. These are (*a*) the expansions of agricultural systems from hearth areas such as Southwest Asia, China, and Mesoamerica, and (*b*) the expansions of the world's major language families. Some of the latter are of course associated with

---

otherwise the identity amounts to little more than a statement that early farming coincides with the present-day distribution of languages" (Blench, forthcoming). I would not disagree on the validity of this type of argumentation, even if it is not completely bullet-proof, given the possibilities of cultural devolution (loss of farming) and late and unnoted diffusion of farming-related vocabulary giving false impressions of the existence of agriculture in early times. It should also be kept in mind that evidence in terms of reconstructed vocabulary is not always available or even immediately attainable, since thorough reconstruction for most language families of the world is still lacking, as is, in many cases, the requisite, basic documentary information. Finally, and perhaps most importantly, it is not the case that the participation of speakers of a certain language family in a neolithic revolution and the subsequent effects on the spread of the language family necessarily have to occur at the level of the proto-language itself. Without doubt, the beginning of Niger-Congo, for instance, would be much older than the introduction of agriculture among prehistoric speakers of languages of this family. So the presence of terms for agriculture in the proto-language of a given family may sometimes not be expected, even if it is clear that speakers of early stages of the language family participated in a neolithic revolution and that this has been consequential for the subsequent expansion of the family.

> predominantly hunter-gatherer populations, but the majority occur in agricultural latitudes and their component languages are spoken by people who were already agriculturalists at the dawn of history**.** Many of these widespread agriculturalist language families, such as Austronesian, Indo-European, Niger-Congo, Uto-Aztecan, and Afroasiatic, had reached their precolonial geographical limits (give or take a few hundred kilometers) long before the local existence of any written records--their spreads belong among prehistoric farmers/pastoralists and small-scale social formations, rather than among the great conquest empires and charismatic world religions of history. Could the early dispersals of agriculture and the early spreads of certain major language families be linked effects of the same underlying set of causes? Do these causes relate to the demographic growth and rapid expansion profiles of early farmers? (Bellwood 2001: 182).

Campbell (2003) has pointed to the problem that the language-farming dispersal model fails to explain why certain major language families do and others do not correlate with prehistoric agricultural areas. To cite a prominent example, why does Indo-European fit the model, but not Uralic? Campbell (2003: 50) gives the following list of language families that are "significantly spread" but do not have agriculture, comparing them to a list of language families where the prediction works better: Tungusic, Uralic, Eskimo-Aleut, Pama-Nyungan, Salishan, Uto-Aztecan, Athabaskan, Algonquian, Siouan, Yuman, Chon, Jê. In his afterthoughts to the volume wherein Campbell (2003) is included, Bellwood defensively remarks that "[t]he immensity and complexity of the human past will always allow other hypotheses to exist, as it will also allow the existence of situations within which the hypothesis manifestly does not work. Critics of the hypothesis will always be able to rub their hands with glee as yet another non-matching situation is hauled out of the annals of archaeology or anthropology and paraded before an awed audience of non-believers" (Bellwood 2003b: 468). I would agree that finding some counterexamples is not enough to demolish a theory. On the other hand, the counterexamples can become so numerous that the theory reduces to just one out of several explanations suited to explain individual situations. This is a problem that cannot simply be done away with by rhetorical means. In the following I shall therefore suggest a modification of the theory which increases its potential for making precise predictions regarding the correlation of language families and prehistoric subsistence patterns on a world-wide scale.

2. Steps Towards a Possible Improvement of the Correlation

The novel suggestion of this paper is that what I call language family ‘density’ is a better

predictor of prehistoric subsistence strategies than the sheer size of a family measured in terms of number of languages or geographical spread.[3] Density is the relationship between the number of languages in a family and the internal linguistic differentiation of the family, as measured in differences in vocabulary. The idea comes from looking at the kinds of language families that represent a problem for the Bellwood/Renfrew hypothesis. Some of the exceptions to the hypothesis are constituted by widespread hunter-gatherer language families. These are usually characterized by a relatively small number of languages; on the other hand, the linguistic differences among the languages are relatively great. Another group of exceptions are some small families spoken by people who have been farmers as long as farming has existed in their area. In such families the internal differentiation is usually small. An example of the latter kind of language family would be the Mayan family of Mesoamerica, which, on a conservative count, consists of 31 languages. Compared with some of the big language families in the world, 31 languages is not a great number, but it is a great number relative to the differences among the languages, which is not very great. Why language families of hunter-gatherers should have a small density (in the sense just mentioned) as compared to those of farmers may be explained as follows. We can imagine, and it has been fact been argued by Daniel Nettle (1999a,b), that the rate of language change will be greater in a small community than in a large one. In a little band of hunter-gatherers it should be easier for individual innovations to perpetuate throughout the whole community than in a larger clusters of village inhabited by sedentary farmers. On the other hand, over time the population expansion will be greater among farmers (Golson 1982), causing a slow spread of the population over an increasingly large area. This, in turn, will lead to dialect differences and eventually the emergence of new languages. It is not necessarily expected, however, that the expansion of a family of languages spoken by farmers must result in families of the size of, say, Austronesian, because its expansion may be impeded by physical or ecological barriers. This is the case with the languages families of Mesoamerica, for instance. The geographical north-south orientation of the land mass connecting the subcontinents prevents a horizontal spread within same ecological zones (Crosby 1986: 18). Additionally, in a situation where farming is taken up roughly simultaneously among neighbouring groups, the competition of equally thriving agriculturalists speaking other languages may represent an impediment to expansion.

In order to transform my scenario for the formation of a language family typical of farmers into something quantifiable I shall propose what I call a *density ratio*. This is obtained

---

3 Geographical spread is a parameter which I do not discuss further in this paper. Obviously the vast geographical extensions of hunter-gatherer families like Uralic, Eskimo-Aleut, and Na-Dene militate against any attempt to correlate geographical spread with agriculture.

by dividing the number of languages, N, in a family with a measure of linguistic diversity, mc.

$$D = N/mc$$

Nobody has so far been able to produce a quantifiable measure of grammatical divergence among languages, but we do have a measure of lexical divergence that might be used for the purpose of calculating the ratio of density. This measure of lexical divergence is provided by the so-called glottochronological method. It is a method developed by Morris Swadesh and others in the early 1950's (Swadesh 1950, 1952, 1967) and is designed to measure the age of a given language family or subgroup on the basis of a formula where the only variable is the degree of differences in vocabulary on a list of 100 or 200 basic meaning items. It is assumed for a fixed time span (such as 1000 years), the number of cognates shared between any two pairs of related languages will be reduced by a constant percentage (Lees 1953). The method has contested for a number of different reasons, and is in the minds of the majority of historical linguists, discredited (e.g., Rea 1958, Fodor 1961, Bergsland and Vogt 1962, Dixon 1997: 35-37). Others believe that the method is, indeed, valid for making archaeological and historical linguistic correlations and apply it in its original form or attempt to refine it. Starostin (2000) is the presently the best known and most widely applied refinement of the original method. I have some reservations about the theory myself. One problem is that its presupposition of a constant rate of changes is in conflict with the aforementioned hypothesis that the size of the speech community will affect rates of change. But my point in using glottochronological data is not to determine absolute chronologies or even relative ones. What I am interested in is only the variable figures that enter into the formula and which give a convenient measure of lexical differentiation.

If we divide the number of languages with the glottochronological time depth ('mc' stands for 'minimal centuries') representing the degree of lexical differentiation, we will obtain a density ratio of the sort that I am looking for.

A relatively large density should correlate with agriculture, while a relatively low density should correlate with hunter-gatherers. In section 5 below I shall compare the success of the predictions of this models with the succes of predictions simply based on the number of languages in the various families.

3. Concerning the correlation procedure

Our aim is to see whether language family sizes or language family densities provide the better correlate with agriculture. This amounts to seeing whether there is some cut-off point in, respectively, size-ranked and density-ranked lists of language families below which there are

correlations with the absence of agriculture and above which there are correlations with the presence of agriculture.

[PLACE FIGURES 1-2 AROUND HERE]

Fig. 1 shows a curve representing the size-ranking of the language families in the data of Table 1 below. As can be seen, the data line up in a way so as to closely approximate a curve of the general shape $y = bx^{-a}$ (an observation which also holds when the data is extended to all of the world's language families). Such a distribution is called a 'power-law distribution' and characterizes many phenomena in the physical, biological, and social worlds (Bak 1996: 12-27). Scholars agree that power-law distributions are ultimately due to stochastic processes although no consensus has been reached as to how they are best explained. A special instance of the distribution, where $y \sim x^{-1}$, is known as Zipf's law, named after George Kingsley Zipf, who first observed that absolute word frequencies are inversely proportional to their rank (Zipf 1949). The discovery that language family sizes have a power-law distribution has interesting implications which, however, exceed the scope of this paper (see Wichmann n.d.). In the present context the usefulness of the observation is that there are very many small families, some intermediate, and only a few large ones and that the distribution is such that they approximately align on a straight line in a log-log plot given that $y = bx^{-a}$ is equivalent to $\log(y) = -a \log(x) + \log(b)$, cf. figure 2. The densities similarly follow a power-law distribution. For the purpose of inserting cut-off points in the distributions, then, it is practical to convert the rankings of family sizes and densities to logarithmic scales. Moreover, I shall calibrate both to a scale running from 1 to 100 in order to render the two distributions more easily comparable. In the case of the language families the calibration requires us to find the values of q and r in the formula

$$N' = q \ln N + r, \quad \text{where } N = \text{language family size and } N' \text{ the calibrated size}$$

and then apply this formula in the calibration. The values of q and r are found by solving the following two equations (where N_{max} is represented by Niger-Congo and N_{min} by Yeniseian, inter alia):

$$N'_{max} = 100 = q \ln N_{max} + r = q \ln 1489 + r$$
$$N'_{min} = 1 = q \ln N_{min} + r = q \ln 2 + r$$

The rounded off values of q and r found by this means are inserted into the calibration formula to yield

$$N' = 14.971 \ln N - 9.377$$

By a similar procedure we arrive at a calibration formula for the densities (derived from the figures $D_{max} = 36.057$ represented by Austronesian and $D_{min} = 0.094$ represented by the mean density of the span estimated for Plateau Penutian).

$$D' = 16.640 \ln D + 40.344$$

4. Data

The language families represented in Table 1 below are only those for which I have had access to glottochronological estimates. Unfortunately the authors who have produced these estimates never provide the data and formulas used. Thus is it difficult to gauge the potential variability in the estimates. Nevertheless, I do not expect the overall picture to vary significantly even if some of the glottochronological figures should be revised. As for the numbers of languages they are taken from *Ethnologue* (Grimes 2000) whenever possible. The reason for this choice is not that I necessarily agree with all the counts, but rather that *Ethnologue* at least seems to apply consistent criteria in its estimates of languages per family. It generally includes entities in the count that more conservative sources would treat as dialects. Nevertheless, this is only an advantage in the present context since a more fine-grained count statistically decreases the effects of subtracting or adding a few units in the case of smaller families. A source such as Ruhlen (1987) often gives radically different figures, which are mostly smaller. I have also calculated density measures using Ruhlen and found that the same six language families that have the highest densities using the *Ethnologue* figures are identical with the ones arrived at using Ruhlen: Austronesian, Niger-Congo, Trans-New Guinea, Sino-Tibetan, Afro-Asiatic, Indo-European, Tai-Kadai (cited in the rank-ordering produced by the data in Ruhlen). For lower densities the two sources yield somewhat different results--to the extent that they are at all comparable. Ruhlen lacks Xincan and Yeniseian and does not have an entity directly corresponding to Moseten-Chon. For the following reasons I have chosen not to base my observations on the data of Ruhlen: the data are older than those of *Ethnologue*, the estimates seem to be more loose, the coarse-grained nature of the figures increases the statistical effects of errors at the level of smaller language families, and, as also already mentioned, some small families are lacking.

Table 1 presents the data that I have been able to accumulate to date which are relevant to the correlation procedure sketched above.

[PLACE TABLE 1 ABOUT HERE]

5. Testing the correlations of farming and language family sizes

In the following I shall compare the success of the prediction that a high number of languages correlates with farming and a low number with its absence with the success of the prediction that a high density correlates with farming and a low number with its absence.

There is a special case, namely that of Australian, which is an embarrassment to both theories. Australian both has a high number of languages (N' = 73.8) and a high density (D' = 57.0). In either correlation strategy Australian will figure among the language families for which we would predict that agriculture is present, which, of course is not the case. One may question the validity of Australian as a genetic unit. However, the problem then carries over to Pama-Nyungan. Pama-Nyungan is either the largest subgroup of Australian or, if one does not accept Australian as a genetic unit, the largest language family in Australia. Given that Pama-Nyungan comprises all the extant Australian languages except those to the far north, both the size and density of this group are very large and constitute just as much a problem for both theories as Australian. One way to escape from the problem is to side with Dixon (2001), who has claimed that Pama-Nyungan is not a genetic unit, but essentially just a Sprachbund, where similarities are due to diffusion. This claim is very controversial, however, not the least among well nigh all the other Australianists actively working in the field of historical linguistics (see papers in Bowern and Koch 2004). I would prefer not to take sides in the issue, although I do admit that the evidence that I have seen in the favor of Pama-Nyungan as a genetic unit (as presented at conferences as well as in personal communication) does seem rather convincing. Australian (and/or Pama-Nyungan) could well constitute an exception to any attempt to correlate prehistoric subsistence patterns with features of language families. So far I do not have an explanation. In the tables below Australian has been left out, but it should kept in mind as a problem to be dealt with in future research.

In the left part of Table 2 the language families are ranked according to (calibrated) size and in the right part according to (calibrated) densities. The darkened areas in each side of the table represent the spans of calibrated figures within which the correlation works without exceptions.

[PLACE TABLE 2 ABOUT HERE]

In the next section we shall discuss some matters of detail, but for the present we can draw the straightforward conclusion that language family densities provide a somewhat better correlation

measure than language family sizes.

Within the span $70 \prec N' \preceq 100$ we can predict the presence of agriculture and within the span $N' \preceq 1$ its absence. Seven families conform to the predictions. For the remaining 29 families no predictions can be made. If Nilo-Saharan is allowed to figure as a special exception, the picture improves. Then we might say that within the span $54 \preceq N' \preceq 100$ the presence of agriculture is expected (with one exception only). All in all, in 19% of all cases the prediction works without exceptions. If we allow for one exception, it works in 35% of all cases.

If predictions are made on the basis of (calibrated) densities, the presence of agriculture correlates with the span $47 \preceq D' \preceq 100$ and its absence with the span $0 \prec D' \preceq 7$. The prediction works without exceptions for 16 cases, while in 20 cases no predictions can be made. The prediction, then, works without exceptions in 43% of all cases.

6. Discussion

With respect to the handling of data, the preceding sections are somewhat crude. One obviously does not deal adequately with 10,000 or more years of the entire human prehistory by plotting some figures in a diagram and correlating them with plus/minus values for an inherently complex phenomenon like prehistoric agriculture. Even if figures and statistics often have the effect of making scholars in the humanities somewhat uneasy I would maintain that they provide a better way to make one's assumptions explicit than vague generalizations formulated in a discursive manner. On the other hand, one should of course bear in mind that figures may sometimes be 'too precise', given that some phenomena are not so easily quantified or described in terms of plus-minus values. I have already commented on some of the problems relating to the N (family size) and D (density) values. The correlation exercise should probably be repeated with another (consistent) count of language family sizes and it would also be preferable to have a more consistent set of D values. Moreover, it may be possible to expand the set of language families when more results from the comparative linguistics of linguistically lesser-known areas such as New Guinea become available. Given these drawbacks the present work may be considered a pilot study only and future research is clearly needed. Leaving these problems aside, I would like to comment briefly on some of the assignments of p(resence) vs. a(bsence) values of the agriculture parameter, moving down the lists in Table 2.

Trans-New Guinea. When first proposed the so-called Trans-New Guinea 'phylum' was somewhat controversial, but now it appears that most experts recognize that there is indeed a very large family concentrated in the central New Guinea highlands, which comprises, if not all the languages that were originally thought to belong to it, at least a great many, that is close to 300 in some reports or 552 in the *Ethnologue* count. It is likely that New Guinea could be a good

laboratory for the kind of correlation that I am interested in establishing. Foley (1986: 277) observes that the central highlands, where the largest family is concentrated, show evidence for agriculture as early as about 7,000 BP (Golson 1977). If we had available glottochronological figures for all of the Papuan families, some of the figures would enable us to make predictions concerning the extent of prehistoric farming in New Guinea. And with more archaeological data the predictions could then be tested.

Nilo-Saharan. For the success of the correlation that departs from sheer language family sizes, the case of Nilo-Saharan is of special importance. I have assigned to this family an "a(bsence)" value for the agriculture parameter. According to Peter Peregrine (personal communication) there is very little evidence of farming during the East African Neolithic (5000-1250 BP), which is the archaeological period that seems to best correlate with early stages of the development of Nilo-Saharan. Domestic cattle herding was central to the economy, with sheep and goats also present. The presence of herding and animal domestication in general may perhaps explain why Nilo-Saharan is large, but if one wants to insist that intensive agriculture is the primary prerequisite for the development of large language families Nilo-Saharan is a problem.

North and South Caucasian. I have assigned "?" values to North and South Caucasian. Again I have consulted with Peter Peregrine, who informs me that these two families correspond to the Caucasian Neolithic in the *Outline of Archaeological Traditions* of the Human Relations Area Files[4]. The Caucasian Neolithic dates to ca. 8000-6500 BP. Although some farming gradually emerged it appears to have been of minor importance. However, there is evidence of the domestication of sheep, pig, and cow. Interestingly, there is also evidence for continued hunting and gathering throughout the period. It seems to be the case, then, that some sort of symbiosis of farmers and hunter-gatherers developed. My correlation procedure predicts that North Caucasian relates to farmers, while South Caucasian relates to hunter-gatherers, providing an intriguing parallel to the archaeological evidence. It would be interesting to take a more detailed look at this situation in future work.

Na-Dene and Algic. The density measure for Na-Dene, Algic, as well as Nilo-Saharan, are almost identical (falling within the range 45.0-45.2), and the three families are just below the cut-off value D' = 47 that predicts the presence of agriculture. Even if the figures fall below the cut-off point–as they should given the absence of farming–it is perhaps nevertheless surprising that the scores are so high. Prehistoric agriculture is only sparsely documented in North America and is largely restricted to the eastern woodlands, where, according to Smith (1992), an agricultural development began around 2000 BC. The high scores of Na-Dene and Algic,

---

4 For online access to Human Relations Area Files cf. http://www.yale.edu/hraf/.

however, suggests that fishing could have an impact on the density of language families almost as great as that of agriculture. Driver (1969: 88), referring to work by Rostlund (1952), notes that "fishing was more productive per acre than hunting or wild plant gathering. It was second only to agriculture in this respect. The relatively sedentary way of life on the Northwest Coast was made possible by the abundance of food available within a small territory." For Na-Dene Gruhn (1997), citing Jacobsen (1989), points to the possibility of a coastal origin of Na-Dene, which conforms to a hypothesis that its spread relates to a subsistence strategy based on fishing. A similar hypothesis may apply to Algic. The cases of Na-Dene and Algic, together with that of Nilo-Saharan, suggest that successful subsistence strategies other than that of farming may have an effect on language family densities. In general, it may be possible to provide a more nuanced correlation of language family sizes with a variety of different subsistence strategies. For the moment, however, I am content to note that the agriculture/density correlation works well in the extremes of the range of density values and that there is a span where predictions are simply not possible. Future work might show whether approaches to the "twilight zone" might be developed.

Altaic. This is the only language family for which I have assigned a 'p/a' value. Altaic is ambiguous since the southwestern and southeastern extremes reach into the West and Southeast Asian agricultural zones. Given that Altaic belongs to the set of cases where predictions are not possible anyhow, there is no need to go into detail concerning the archaeological correlates of this family. It should remarked, however, that it might be a case where one could refine the correlation by zooming in on particular subgroups, make glottochronological calculations, and try to establish predictions at shallower levels. This is yet another possible item for future research.

7. Conclusion

The purpose of this paper has been to test the language/farming dispersal hypothesis on a global scale using explicit, quantitative measurement rather than loose, discursive statements. It turned out that language family sizes (i.e., the number N of languages contained within a given family) is a parameter which only leads to correct predictions in a few cases–about one fifth of all cases or around one third if one exception is allowed for. A better parameter for the correlation is that of density (D), defined as the number of languages in a family divided by the glottochronological time depth. This parameter allows for correct predictions in 43% of all cases, leaving a little of half of the cases simply nonpredictable. For the glottochronological time depth one might probably substitute any other quantitative measure of internal divergence, but no such alternative measurements are available.

Both the correlations departing from N and and those departing from D break down in the case of language families that have intermediate values. If nothing else, then, this study has shown that there is a level at which one should not try to make predictions using either parameter. One of the cases in point is Uto-Aztecan, which has been much debated as a possible case of farming-cum-language dispersal.

The focus on farmers vs. hunter-gatherers leads to a somewhat reduced picture of the real complexity of prehistory. Obviously other subsistence patterns are attested. Indeed, it was mentioned that fishing and cattle-herding may also relate to language family densities. But these additional subsistence strategies do not provide as clear linguistic fingerprints as do the ones we have been focusing on here.

Some special cases where future research may profitably be carried out were mentioned. They include the Caucasus region and New Guinea. For implementing and testing the correlation procedure proposed these areas, which are among some of the linguistically most diverse on the planet, are very interesting. Both may have witnessed the simultaneous presence of hunter-gatherers and agriculturalists, so the procedure might help us to shed light on the question of which ethnic groups performed which functions within the social ecologies of the regions.

The case of Australian (and/or Pama-Nyungan) was left out of consideration since it uniquely defies correct predictions by either one of the two correlation strategies discussed. Linguistic reconstructions of Australian languages are fast under way, so there is hope that it may soon be resolved whether or not this continent truly presents an exception to our theory or, alternatively, whether it is special in some sense whereby it may be removed from the purview of our theory.

In the context of the present volume of papers it is appropriate to mention that in the case of Indo-European we would strongly predict the presence of agriculture, regardless of whether we use N or D values for the correlation. It is a hotly debated issue whether the spread of Indo-European was simultaneous with or happened later than the spread of farming. I would like to stress that my results do not necessarily bear on this issue. For significant effects of farming to show up in the make-up of a language family it does not seem necessary that the proto-language be exactly contemporaneous with a neolithic revolution. Moreover, a proto-language is just one of many points along a continuum of linguistic evolution; it is arbitrarily singled out by our methods of linguistic reconstruction (Wichmann 2003) and does not necessarily warrant the special status which is automatically conferred upon it be virtue of its attachment to the apparent root of an evolutionary tree.

References

Ammerman, Albert J. and Luigi Luca Cavalli-Sforza. 1984. *The Neolithic Transition and the Genetics of Populations in Europe.* Princeton, NJ: Princeton Univ. Press.

Bak, Per. 1996. *How Nature Works: The Science of Self-Organized Criticality*. New York: Springer-Verlag New York.

Bellwood, Peter. 1994. An archaeologist's view of language macrofamily relationships. *Oceanic Linguistics* 33: 391-406.

Bellwood, Peter. 1997. The prehistoric cultural explanations for the existence of widespread language families. In Patrick McConvell and Nick Evans (eds.), *Archaeology and Linguistics: Aboriginal Australia in Global Perspective*, pp. 23-34. Melbourne: Oxford University Press.

Bellwood, Peter. 2001. Early agriculturalist population diasporas? farming, languages and genes. *Annual Review of Anthropology* 30: 181-207.

Bellwood, Peter. 2003a. Farmers, foragers, languages, genes: the genesis of agricultural societies. In Bellwood, Peter and Colin Renfrew (eds.), pp. 17-28.

Bellwood, Peter. 2003b. Enlightenment or obfuscation? Some afterthoughts. In Bellwood, Peter and Colin Renfrew (eds.), pp. 467-469.

Bellwood, Peter and Colin Renfrew (eds.). 2003. *Examining the Farming/Language Dispersal Hypothesis*. Cambridge: McDonald Institute for Archaeological Research.

Blench, Roger. Forthcoming. Archeology and language: methods and issues. Chapter prepared for *The Blackwell's Companion to Archaeology*, edited by John Bintliff.

Bergsland, K. and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3: 115-153.

Bowern, Claire and Harold James Koch (eds.). 2004. *Australian Languages: Classification and the Comparative Method*. Amsterdam Studies in the Theory and History of Linguistic Science, Series IV: Current Issues in Linguistic Theory. Amsterdam: John Benjamins.

Campbell, Lyle. 1997. *American Indian Languages. The Historical Linguistics of Native America*. Oxford Studies in Anthropological Linguistics. New York-Oxford: Oxford University Press.

Campbell, Lyle. 2003. What Drives Linguistic Diversification and Language Spread? In: Bellwood, Peter and Colin Renfrew (eds.), pp. 49-63.

Crosby, A. 1986. *Ecological Imperialism*. Cambridge: Cambridge University Press.

Dixon, Robert M.W. 1997. *The Rise and Fall of Languages. Cambridge: Cambridge University Press.*

Dixon, R.M.W. 2001. The Australian linguistic area. In: Aikhenvald, Alexandra Y. and Dixon, R. M. W. (Eds.), *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, pp. 64-144. Oxford: Oxford University Press.

Driver, Harold E. 1969. *Indians of North America*. A2nd. rev. ed. Chicago and London: The University of Chicago Press.
Fodor, I. 1961. The validity of glottochronology on the basis of the Slavic languages. *Studie Slavica* 7: 295-346.

Foley, William A. 1986. *The Papuan Languages of New Guinea*. Cambridge: Cambridge University Press.

Fortescue, Michael. 1998. *Language Relations across Bering Strait*. New York and London: Cassell.

Fowler, Catherine S. 1983. Lexical clues to Uto-Aztecan prehistory. *International Journal of American Linguistics* 49: 224-57.

Golson, Jack. 1977. No room at the top: agricultural intensification in the New Guinea highlands. In Jim Allen et al. (eds.), *Sunda and Sahul: Prehistoric Studies in Southeast Asia, Melanesia, and Australia*, pp. 601-638. New York: Academic Press.

Golson, Jack. 1982. The Ipomean revolution revisited: society and the sweet potato in the Upper Wahgo Valley. In Andrew J. Strathern (ed.), *Inequality in the New Guinea Highlands*, pp. 109-136. Cambridge: Cambridge University Press.

Grimes, Barbara F. (ed.) and Joseph E. Grimes (cons. ed.). 2000. *Ethnologue*. Vol. 1: *Languages of the World*. 14th ed. Dallas: SIL International.

Gruhn, Ruth. 1997. Language classification and models of the peopling of the Americas. In: Patrick McConvell and Nicholas Evans (eds.), *Archaeology and Linguistics. Aboriginal Australia in Global Perspective*. Melbourne: Oxford University Press.

Heine-Geldern, R. 1932. Urheimat und früheste Wanderungen der Austronesier. *Anthropos* 27: 543-619.

Hill, Jane. 2001. Proto-Uto-Aztecan: a community of cultivators in Central Mexico? *American Anthropologist* 103.4: 913-934.

Hill, Jane. 2003. Proto-Uto-Aztecan cultivatio and the northern devolution. In Bellwood, Peter and Colin Renfrew (eds.), pp. 331-340.

Jacobsen, William H. Jr. 1989. The Pacific orientation of western North American languages. Paper presented at the Circum-Pacific Prehistory Conference, Seattle, August 1989.
Kaufman, Terrence. 1974. Meso-American Indian languages. In: *Encyclopædia Brittanica* (15th ed.), Vol. 22, pp. 767-774.

Kaufman, Terrence. 1990. Language history in South America. In Doris L. Payne (ed.), *Amazonian Linguistics*, pp. 13-73. Austin: University of Texas Press.

Kaufman, Terrence and Victor Golla. 2000. Language groupings in the New World: Their reliability and usability in cross-disciplinary studies. In Colin Renfrew (ed.), *America Past, America Present: Genes and Languages in the Americas and Beyond*, pp. 47-57. Cambridge: McDonald Institute for Archaeological Research.

Lees, Robert B. 1953. The basis of glottochronology. *Language* 29: 113-127.

Nettle, Daniel. 1999a. Is the rate of linguistic change constant? *Lingua* 108: 119-136.

Nettle, Daniel. 1999b. *Linguistic Diversity*. Oxford and New York: Oxford University Press.

Peiros, Ilia and Victor Shnirelman. 1992. Towards an understanding of proto-Dravidian prehistory. In Vitaly Shevoroshkin (ed.), *Reconstructing Languages and Cultures*, pp. 70-71. Bochum: Brockmeyer.

Price, T. Douglas. 1996. The agricultural frontier and the transition to farming in the circum-Baltic region. In: Harris, David R. (ed.), *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, pp. 346-362. London: UCL Press.

Rea, John A. 1958. Concerning the validity of lexicostatistics. *International Journal of American Linguistics* 24: 145-150.

Renfrew, Colin. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins*. London: Jonathan Cape.

Renfrew, Colin. 1994. World linguistic diversity. *Scientific Amerian* 270: 116-123.

Renfrew, Colin. 2000. At the edge of knowability: towards a prehistory of languages. *Cambridge Archaeological Journal* 10.1: 7-34.

Renfrew, Colin. 2003. 'The emerging synthesis': the archaeogenetics of farming/language dispersals and other spread zones. In: Bellwood, Peter and Colin Renfrew (eds.), pp. 3-16.
Rockmore, Daniel. 2004. Are you my mother . . . tongue? *SFI Bulletin* 19.1: 11-15.

Romney, A.K. 1957. The genetic model and Uto-Aztecan time perspective. *Davidson Journal of Anthropology* 3: 25-41.

Rostlund, Erhard. 1952. *Freshwater Fish and Fishing in Native North America*. University of California Publications in Geography, No. 9.

Ruhlen, Merritt. 1987. *A Guide to the World's Languages*, Vol. 1: *Classification*. London: Edward Arnold.

Smith, Bruce. 1992. Prehistoric plant husbandry in Eastern North America. In: C.Wesley Cowan and Patty Jo Watson (eds.), *The Origins of Agriculture*, pp. 101-120. Washington, D.C.: Smithsonian.

Starostin, George. 2004. Evolution of Khoisan click systems in the light of glottochronological

data. Paper presented at the Prehistoric Chronology Workshop, Santa Fe Institute, March 1-5, 2004.

Starostin, Sergei. 2000. Comparative-historical linguistics and lexicostatistics. In: Colin Renfrew, April McMahon and Larry trask (eds.), *Time Depth in Historical Linguistics*, vol. 1, pp. 223-265. Cambridge: McDonald Institute for Archaeological Research.

Starostin, Sergei. 2004. Indo-European glottochronology and homeland. Paper presented at the Prehistoric Chronology Workshop, Santa Fe Institute, March 1-5, 2004.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16: 157-167.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96: 452-463.

Swadesh, Morris. 1967. Lexicostatistic classification. *Handbook of Middle American Indians*. 5, pp. 100-115. Austin: University of Texas Press.

Wichmann, Søren. n.d. On the power-law distribution of language family sizes. Submitted.

Wichmann, Søren. 1999. On the relationship between Uto-Aztecan and Mixe-Zoquean. *Kansas Working Papers in Linguistics* 24.2: 101-13.

Wichmann, Søren. 2003. Contextualizing proto-languages, homelands and distant genetic relationship: some reflections on the comparative method from a Mesoamerican perspective. In: Bellwood, Peter and Colin Renfrew (eds.), pp. 321-329.

Zipf, George K. 1949. *Human Behavior and the Principle of Least-Effort*. Cambridge: Addison-Wesley.

Zvelebil, Marek and Kamil Vaclav Zvelebil. 1988. Agricultural transition and Indo-European dispersals. *Antiquity* 62: 574-583.

Table 1. Data on language family sizes, densities, and correlation with agriculture

Language family	N	N'	mc	D	D'	Agric.	Source for N	Source for mc
Indo-European	443	81.8	70	6.3	71.0	p	A	B
Altaic	65	53.1	77	0.8	37.5	p/a	A	C
Uralic	38	45.1	60	0.5	32.7	a	A	C
Yeniseian	2	1.0	5	0.4	25.1	a	A	C
North Caucasian	34	43.4	60	0.6	49.8	a	A	C
South Caucasian (Kartv.)	5	14.7	40	0.1	5.7	a	A	C
Chukotka-Kamchatka	5	14.7	40	0.1	5.7	a	A	D
Eskimo-Aleut	11	26.5	30	0.4	23.6	a	A	E
Dravidian	75	55.3	40	1.9	50.8	p	A	C
Niger-Congo	1489	100.0	100	14.9	85.3	p	A	F
Afro-Asiatic	372	79.2	113	3.3	60.2	p	A	G
Nilo-Saharan	199	69.9	150	1.3	45.0	p	A	H
Khoisan	29	41.0	111	0.3	44.7	a	A	I
Hmong-Mien (Miao-Yao)	32	42.5	40	0.8	36.6	p	A	C
Tai-Kadai (Daic)	70	54.2	30	2.3	54.4	p	A	C
Sino-Tibetan	365	79.0	60	6.1	70.4	p	A	J
Austronesian	1262	97.5	35	36.1	100.0	p	A	K
Trans-New Guinea	552	85.1	100	5.5	68.8	p	A	C
Australian	258	73.8	95	2.7	57.0	a	A	C
Na-Dene	47	48.3	35	1.3	45.2	a	A	E
Algic	40	45.8	30	1.3	45.1	a	A	E
Caddoan	5	14.7	33	0.2	8.9	p	O	E
Salishan	27	40.0	45	0.7	31.8	a	A	E
Plateau Penutian	4	11.4	35-50	0.1	1.0	a	O	E
Iroquoian	10	25.1	35	0.3	19.5	p	A	E
Hokan	28	40.5	88	0.3	21.3	a	A	L

Mayan	69	54.0	42	1.6	48.6	p	A	E
Otomanguean	172	67.7	60	2.9	57.9	p	A	E
Uto-Aztekan	62	52.4	48	1.3	44.6	a	A	E
Mixe-Zoquean	16	34.9	36	0.4	26.9	p	A	M
Totonakan	11	26.5	26	0.4	26.0	p	A	M
Xincan	4	11.4	10	0.4	25.1	p	O	M
Barbakóan	7	19.8	33	0.2	14.5	p	A	N
Quechuan	46	47.9	15	3.1	59.0	p	A	N
Moseten-Chon	5	14.7	51	0.1	1.7	a	N	N
Tupían	70	54.2	55	1.3	44.4	p	A	E
Káriban	29	41.0	37	0.8	36.3	p	A	N

Legend: N = number of languages; N' = calibrated number of languages; mc = minimal centuries; D = density; D' = calibrated density; Agric. = agriculture (p = presence, a = absence of agriculture.) Source codes: A = Grimes (2000); B = S. Starostin (2004); C = Rockmore (2004) citing results of S. Starostin and associates; D = Fortescue (1998: 39); E = Kaufman and Golla (2000); F = rough estimate; G = Militarev (2004); H = C. Ehret (pers. comm.); I = G. Starostin (2004); J = Peiros and Shnirelman (1998); K = Foley (2000: 362-63); L = Swadesh (1967); M = Kaufman (1974); N = Kaufman (1990); O = Campbell (1997).

Table 2. The correlation of language family sizes and farming vs. the correlation of language family densities and farming

N'	Language family	Agric.		D'	Language family	Agric.
100.0	Niger-Congo	p		100.0	Austronesian	p
97.5	Austronesian	p		85.3	Niger-Congo	p
85.1	Trans-New Guinea	p		71.0	Indo-European	p
81.8	Indo-European	p		70.4	Sino-Tibetan	p
79.2	Afro-Asiatic	p		68.8	Trans-New Guinea	p
79.0	Sino-Tibetan	p		60.2	Afro-Asiatic	p
69.9	Nilo-Saharan	a		59.0	Quechuan	p
67.7	Otomanguean	p		57.9	Otomanguean	p

55.3	Dravidian	p		54.4	Tai-Kadai (Daic)	p
54.2	Tupían	p		50.8	Dravidian	p
54.2	Tai-Kadai (Daic)	p		49.8	North Caucasian	?
54.0	Mayan	p		48.6	Mayan	p
53.1	Altaic	p/a		45.2	Na-Dene	a
52.4	Uto-Aztekan	a		45.1	Algic	a
48.3	Na-Dene	a		45.0	Nilo-Saharan	a
47.9	Quechuan	p		44.7	Khoisan	a
45.8	Algic	a		44.6	Uto-Aztekan	a
45.1	Uralic	a		44.4	Tupían	p

43.4	North Caucasian	?	37.5	Altaic	p/a
42.5	Hmong-Mien (Miao-Yao)	p	36.6	Hmong-Mien (Miao-Yao)	p
41.0	Khoisan	a	36.3	Káriban	p
41.0	Káriban	p	32.7	Uralic	a
40.5	Hokan	a	31.8	Salishan	a
40.0	Salishan	a	26.9	Mixe-Zoquean	p
34.9	Mixe-Zoquean	p	26.0	Totonakan	p
26.5	Totonakan	p	25.1	Xincan	p
26.5	Eskimo-Aleut	a	25.1	Yeniseian	a
25.1	Iroquoian	p	23.6	Eskimo-Aleut	a

19.8	Barbakóan	p		21.3	Hokan	a
14.7	South Caucasian (Kartv.)	a		19.5	Iroquoian	p
14.7	Moseten-Chon	a		14.5	Barbakóan	p
14.7	Chukotka-Kamchatka	a		8.9	Caddoan	p
14.7	Caddoan	p		5.7	Chukotka-Kamchatka	a
11.4	Plateau Penutian	a		5.7	South Caucasian (Kartv.)	?
11.4	Xincan	p		1.7	Moseten-Chon	a
1.0	Yeniseian	a		1.0	Plateau Penutian	a

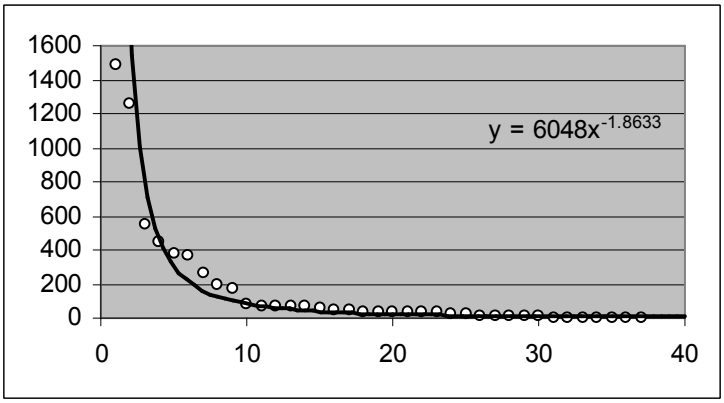

Figure 1. Language family sizes (languages per family) in the data ranked in descending order on the x-axis and plotted against the numbers corresponding to the ranks on the y-axis.

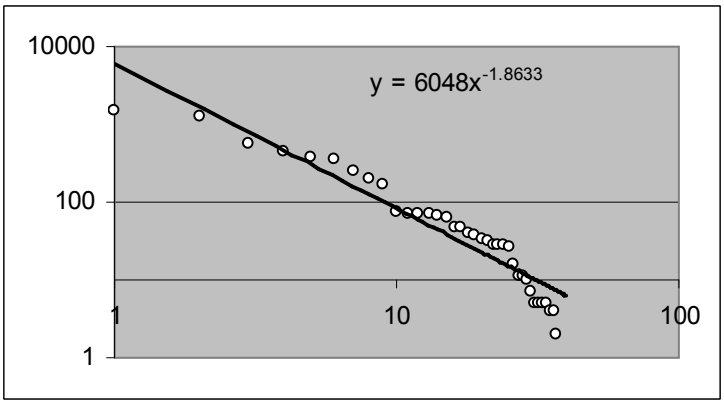

Figure 2. Language family sizes (languages per family) in the data plotted on logarithmic axes.