COMPARATIVE-HISTORICAL LINGUISTICS AND LEXICOSTATISTICS

Sergei Starostin

COMPARATIVE-HISTORICAL LINGUISTICS AND LEXICOSTATISTICS

[This is a translation, done by I. Peiros and N. Evans, of my paper "Sravnitel'no-istoričeskoe jazykoznanie i leksikostatistika", in "Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka", Moscow 1989. I have introduced, however, a number of modifications into the final English text — basically rewritten it again, since the English version needs English examples and etymologies, not Russian ones.]

The last two decades have witnessed a fundamental advance in the techniques of comparative linguistic research. A prolonged period of comparative work with a wide range of language families has laid the foundation for the study of genetic relationships between remotely related languages or language groups. The first step in this direction was taken by V.M. Illich-Svitych in his seminal work 'Towards a comparison of the Nostratic languages' in which, with a combination of rigorous methods and intuitive flare, he begins to demonstrate the relatedness of a number of languages of the Old World.

This new level of comparative studies appears completely legitimate. In fact, if we take the theory of language divergence as axiomatic, we have to concede the fact that from around the sixth millenium B.C. to the first millenium B.C. there was quite a number of different reconstructable proto-languages throughout the world. Once the level of reconstruction of various proto-languages is improved, the question inevitably arises: are any of these proto-languages genetically related and, if so, can we prove this relationship?

To the first part of this question we must now answer in the affirmative. In effect, the absence of genetic relationships between proto-languages (e.g. unrecorded prehistoric languages) could only be explained by the independent arisal of these languages in different parts of the world at approximately the same rather recent period — which would contradict not only simple common sense, but also all our ideas about the history of the human kind and its language. Therefore, at least some existing language families should turn out to be related at an earlier level.

The second part of this question is much more complicated. What is the method of establishing and proving ancient genetic relationships between language families? To this, evidently, there can be only one answer: the classical method of comparative historical linguistics, that is the discovery of a system of regular sound correspondences between proto-languages that is valid for the majority of lexical and morphological items, and the reconstruction of an earlier system based on those correspondences. It is precisely this method that was used by Illich-Svitych in his reconstruction of Common Nostratic.

However, opponents of remote comparison point out that if two languages develop independently for about ten thousand years they inevitably replace practically the entire original lexicon[1]. But ten thousand years or so is exactly the time depth proposed for most macrofamilies. So what can we actually compare in such a case?

Now the situation here is indeed quite complicated, and the probability of establishing the genetic affiliations of language isolates such as Basque, Sumerian or Ainu is extremely small, precisely because of

[1] In fact this is not entirely the case. The theory of glottochronology maintains that the original 100-word list of the basic lexicon will be entirely replaced in a language over a period of about 15-20 thousand years. If one considers that, in addition to this, any change in the meaning of the word, however small, is regarded by glottochronological theory as the loss of that word, then it becomes clear that the list of roots in the language may be very conservative and that basic roots (perhaps with some semantic modification) can be maintained much longer than words. This motivates my method of root glottochronology, to be discussed below.

their early separation from any more inclusive language group, and subsequent independent development. But in most cases the situation is different: we have language families and thus the possibility, via the reconstructed proto-languages, of comparing intermediate stages rather than modern languages. Just as we accept that the degree of resemblance between reconstructed proto-Germanic and proto-Slavonic is greater than between modem German and modern Russian, so we must likewise agree that, if proto-Indo-European and proto-Kartvelian are related, then the resemblance between them should be greater than that between modern Russian and Georgian. So the solution, and at the same time the only possible way of dealing with remotely related languages, is in the so-called stepwise reconstruction and comparison of intermediate proto-languages, and the reconstruction from them of an ancestral proto-language.

These are general considerations. In practice, comparativists who study remote relationships face an additional and more serious danger: the possibility of equating genetically unrelated items. These cases can be divided into several categories:

(a) accidental resemblances, such as English *woman* and Old Japanese *womina* 'woman'.

(b) ideophones, such as Russian *kukushka* and English *cuckoo.*

(c) loans, e.g. proto-Japanese **kui* and proto-Austronesian **kaju'* 'tree, wood'.

Not a few works attempting to prove genetic relationships between different language families have foundered just because they failed to take accidental resemblances and old loans into account[2]. The possibility of this kind of mistake is especially great in the case of so-called 'binary' comparisons (which have been particularly popular in American linguistics) — that is, the comparison of two remotely-related languages or proto-languages. In fact, if we find in Indo-European a root **per* 'front side' and in Dravidian a root **pir-* 'back side', this may well be an accidental resemblance, whose probability could be estimated statistically. But if we add to this pair Altaic **p'era* 'lower side', Uralic **pera* 'back part' and Kartvelian **ṗir-* 'front side' (see: Ilich-Svitych 1971:27) then the probability of coincidental resemblance between all these stems diminishes. If we can find a large number of sets of this kind, demonstrate the regular character of phonological correspondences within them, and discover morphological parallels, then the genetic relationship could be regarded as proven[3].

Even if we agree that the classical comparative model could be used for the study of remote linguistic relationships, there still remains a host of issues which have not received a proper treatment, but which may be of crucial importance: in particular, the closely interrelated problems of language classification and of dating linguistic divergence.

So — if the Nostratic family exists, then:

(1) What are its branches? For example, are there reasons to divide it into Eastern and Western Nostratic branches? Do Turkic, Mongolian and Tungusic form separate branches of Nostratic, or should they be grouped into a single Altaic branch?

(2) What are its limits? There are some doubts about the correctness of including the Afroasiatic languages in Nostratic, and it is possible that these languages form another macrofamily of approximately

[2] Critics of remote comparisons sometimes pay too much attention to the second type of case — sound symbolism — as a factor preventing the comparison of remotely related languages. This is the so-called theory of 'elementary relationship'. In my view, however, the role of sound symbolism is exaggerated. It is unclear, for example, why one should think that the deictic stems like **mV* 'first person pronoun' or **ṭV* 'demonstrative pronoun' should be regarded as examples of sound symbolism. Therefore I will not discuss this theory in detail.

[3] All the above remarks naturally presuppose the classical conception of a 'genetic tree'. If we do not agree with the reality of 'proto-languages', i.e. if we do not think that the reconstructed systems correlate with or are in some sense similar to real languages, then the problem of comparing proto-languages does not arise. Arguments about the problem of remote relationships of languages thus depend on whether one agrees or not with the axioms of historical linguistics, and are thus beyond the scope of this publication.

the same time depth as Nostratic proper. On the other hand, in some publications on Nostratic there are suggestions that the proto-Nostratic, as reconstructed by Illich-Svitych, could be the proto-language of humankind. How can we prove or reject such statements?

(3) What approximate dates are involved? For example, how could we estimate the time-depth of the different branches of Nostratic? In the literature various estimates for proto-Nostratic range from the ninth to the twelfth millenium B.C or even earlier.

The same questions, of course, could be asked not only about Nostratic but about any other putative macro-family. Moreover, on close examination it turns out that traditional historical linguistics often gives us contradictory answers even in the case of much younger language families, whose existence is beyond doubt. For example, the limits of the Indo-European family can be regarded as settled, since it is most unlikely that some other known languages could be included in Indo-European, and at the same time it is unlikely that some languages that are regarded as Indo-European do not in fact belong to this family. But the problems of its internal subgrouping and its time depth, as well as its homeland, are still widely debated.

In many cases it is obvious that languages within a family are not equally closely related: one can find many features common to all Slavic languages but not to, for example, Germanic, and these features allow us to assume that Russian is closer to Bulgarian than to Swedish. Evidently if we could measure these distances with the same metric, then we could:

(a) create internal classifications of any language family (or at least evaluate existing classifications) and thus answer questions about internal subgrouping.

(b) locate a language or language family within higher-level units such as families or macro-families and thus answer questions about the limits of families.

(c) combine this metric with a time-sca1e to estimate the probable time- depth of linguistic divergence, thus answering questions about chronology.

It is thus of central importance to find quantifiable characteristics of languages that can be used as objective criteria in establishing distances between languages[4]. Measurements based on phonological, morphological or syntactic similarities obviously can not be used as such criteria, since such systems are known to change rapidly and radically in some cases, and to be extremely conservative in others. The rate of change in these areas is thus uneven and in any case it is not obvious that it could be measured: a statistical measurement of their rate of change is hindered by the small number of elements (phonemes, grammatical morphemes), which prevents us from obtaining sets that are sufficiently populous for statistical tests[5].

The one domain of language in which the rate of change appears to be even (see on p. below) and which is at the same time an appropriate object for quantitative statistical measures is the lexicon. Therefore I would like to discuss, in the remainder of this paper, various aspects of lexico-statistics and glottochronology.

Despite a number of critical works showing many inadequacies of the glottochronological method invented by Swadesh in the 1950s (to the point where it seemed at one time that glottochronology had been fully discredited), I think it would be not only premature to abandon it, but also incorrect. Every comparativist who has worked with glottochronology knows that closely related dialects usually have a cognacy rate of 90% or more on Swadesh's l00-word list: closely related languages (such as those within the

[4] It is obvious that such criteria could also be used to verify the relationship between languages, and thus the distance between two unrelated languages should be estimated as infinite.

[5] Of course it does not follow from this that phonological and morphological data cannot be used in the genetic classification of languages, since it is well-known that most existing classifications are based on such data. I am only stating that such data are insufficient for quantitative evaluations of the distances between languages. Thus we can separate the Min dialects from all other Chinese dialects by the fact that they maintain the oppositions **b — *bh, *d — *dh* etc., which are lost in all other dialects. But this feature does not help us estimate the depth of their separation.

Slavic, Romance, Germanic or Turkic groups — that is those which diverged around one and a half to two thousand years ago) share from 70 to 80% of items on this list, and language families such as Indo-European which split up five or six thousand years ago have a rate of 25 to 30%. Once we start to talk about more ancient families such as Uralic or Altaic we find a rate of at most 10 or 20 percent. Finally, the cognacy rate for modern languages belonging to different branches of such a macro-family as Nostratic is even less — around 5 — 9 %.

These figures are, to be sure, approximate, and more precise data will be given below. But one should note that at each level the cognacy rate is systematically replicated: no Slavic language has a cognate rate with any other Slavonic language of less than 75%; no Indo-European language has a rate of more than 35% or less than 20% with any other Indo-European language of a different branch. It would be quite absurd to obtain a figure, say, of 50% between Russian and Polish, or Russian and German. It is well-known that similar rates of cognacy are found between languages related at an equivalent level in other language families, such as Austronesian, Uralic, Sino-Tibetan etc. All these considerations give us some indication that the rate of change in the lexicon (in some of its domains at least) really might be steady and universal. There is thus reason to discuss once again the fundamental postulates and methodologies of glottochronology.

The fundamental principles of Swadesh's glottochronology could be represented by five postulates, well-formulated in Arapov, Herz (1974):

"1. In the lexicon of any language one can distinguish a particular portion which we will call basic or stable.

2. One can provide a list of meanings which in any language of the world will be represented by words from its basic vocabulary...We shall say that these words form the basic list BL. Let us represent the number of words in BL by N_o.

3. The proportion *P* of words from BL which remain (i.e. are not replaced by other words) over a time interval *t* ... is constant. That is, it depends only on the amount of time elapsed, and not on how that interval was chosen, or on which words from what language are considered.

4. All words from BL are equally likely to be retained or not to be retained during a particular period of time.

5. The probability of a word from a proto-language's BL being retained in the BL of one of its daughter languages is independent of its probability of being maintained in the comparable list of any other daughter language." (ibid: 21-5)

All the above postulates are used to obtain the basic mathematical equation of glottochronology:

$$(1)\ N(t) = N_o e^{-\lambda t}$$

where the time elapsed between two points of development is denoted by *t* and is measured in millenia; N_o is the initial BL, λ is the 'rate of loss' of words from N_o; and *N(t)* is the proportion of words from the initial list retained at time *t*. Thus, if the original list includes one hundred words and is regarded as 1, and if $\lambda = 0.14$ (i.e. during one millenium fourteen out of one hundred words will be lost), and time elapsed from the beginning of dispersion is, for example, two thousand years, then

$$N(2) = e^{-0.14*2} = 0.76 \text{ (i.e. seventy six words).}$$

Knowing the proportion of words from BL retained in a given language we can calculate, with the help of logarithms the period of elapsed time:

$$(2)\ t = \frac{\ln N(t)}{-\lambda \times N}$$

According to the fifth postulate we assume that the development of two languages from one proto-language was independent. Thus, knowing the cognacy rate in BL lists in two or more related languages, we can work out the time since their separation. The retention rate between n languages, which have a common ancestor with a single BL, is:

$$(3)\quad N_n(t) = N_o e^{-n\lambda t}$$

and the time of their separation can he obtained by the formula

$$(4)\quad t = \frac{\ln N_o(t)}{-n\lambda N_o}$$

Thus, if λ = 0.14 as above, and the retention rate between two languages is 0.8 (i.e. 80%), then the time since separation is

$$t = \frac{\ln 0.8}{-2\times 0.14} = 0.8$$ (i.e. about 800 years)[6]

One should note in particular that all datings so obtained are probabilistic rather than absolute in character, and allow the possibility of errors of various magnitudes, whose confidence interval may theoretically be calculated, but a discussion of this is beyond the scope of this article.

The 'rate' λ is a constant which has been established empirically on the basis of samples from languages with a long history recorded over the last millenium or more. The value of 0.14 for λ which was used above is not arbitrary: precisely this constant was postulated by the founding figure of glottochronology, Morris Swadesh, for his one hundred word BL (see, e.g., Swadesh 1960).

The above theory of glottochronology (whose mathematical apparatus is practically borrowed from the physical theory of radioactive split) is rather elegant and simple. Unfortunately, however, we must hasten to point out that for linguistic datings, Swadesh's version of glottochronology is inappropriate[7]. In practice, in all cases of historically recorded linguistic events we find one and the same outcome: all dates given by standard glottochronology are much younger than the historical records suggest. Let us try to discuss the reasons for this phenomenon.

First of all we should discuss the question of whether the retention rate of BL is in fact constant

[6] In publications on glottochronology one often meets a somewhat simplified notation of the mathematical correlations discussed here, where instead of 'rate' one uses the coefficient of lexical retention *$r = 1-\lambda$*. The following notational variants of the formula are frequently encountered:

(l) as (l') $c = r^t$ where c corresponds to N(t), and N_o is assumed to be 1.

(2) as (2') $t = \frac{\log c}{\log r}$

(3) as (3') $c = r^{nt}$, or more often as $c = r^{2t}$ where n = 2.

(4) as (4') $t = \frac{\log c}{n \log r}$, again more often encountered as

$t = \frac{\log c}{2 \log r}$, for n = 2.

[7] Perhaps this is the reason why many scholars using the glottochronological method no longer try to use it for absolute datings, but only for relative datings (i.e. for creating genetic trees).

(recall that this is the third postulate of glottochronology). This question has been the subject of wide debate, and it is obvious that in the absence of a constant rate the whole procedure of glottochronology becomes senseless.

In the literature on glottochronology one can find the claim that 'the assumption about changes in the lexicon does not apply to languages with an established literary tradition' (see Yakhontov 1984:45). This position held by supporters of glottochronology is a reaction to criticisms put forward in the classic article by Bergsland & Vogt (1962). These scholars, in discussing Scandinavian material, have shown that the rate of change in the BL for Icelandic over the last millenium was only about 0.04 (retention rate *r* = 0.96) while in literary Norwegian (Riksmal) it was about 0.2 (*r* = 0.8). Accordingly we obtain, using Swadesh's value of 0.14 for the constant, improbable results: from 100 to 150 years for Icelandic and 1400 years for Riksmal, despite the fact that both languages developed from the same source, independently, over about 1000 years. Comparable results were obtained by O'Neil (1964) from a comparison of Icelandic and Faroese BLs: the dating of their divergence is known historically (C10 A.D.), but the 94% of shared lexicon between these languages suggests, according to Swadesh's formula, that the divergence started 200 years ago.

These are, it would seem, obvious facts, and they forced scholars to make allowances for the effects of a literary norm which hinders the development of the lexicon — though in the case of Riksmal, as we have seen, we have on the contrary a more rapid development. Let us try, however, to discuss the Scandinavian case.

It is easier to explain the accelerated development of Riksmal. This language is actually a hybrid of Norwegian and Danish, and its one hundred word list includes eleven Danish, three Swedish and two German loans. If we treat all these loans as replacements in our comparison with Old Icelandic then we will get a rapid rate of development for Norwegian. On the other hand, if we exclude these borrowings from consideration we get a rate for Icelandic equal to 0.05, with six words having changed: *eta* 'eat' > *borða, lauf* 'leaf' > *blað* , *verr* 'man' > *maður, varmr* 'warm' > *hlyr, mani* 'moon' > *tungl*. For Riksmal the rate is 0.05, with four words replaced — *eldr* 'fire' > *varme, lauf* 'leaf' > *blad, hold* 'meat' > *kjött* and *sa* 'that' > *den* — from the list of 84 original words. A very similar rate will be obtained for other Scandinavian languages: 0.05 for Faroese, and 0.04 for Swedish, Danish and the Norwegian dialects Gwestal and Sandnes[8].

Thus we can see that if we regard borrowings as lexical replacement we get serious errors. This leads to an important conclusion: before doing glottochronological calculations one should eliminate all borrowings from the BL list (at least where there have been intensive borrowings between two neighbouring languages), paying attention only to the rate of change within non-borrowed lexicon[9].

If we do so, the rate of development of the BL lexicon in Scandinavian languages ends up being stable and not significantly different from 0.05. It is clear, however, that this speed is much slower than the

[8] These values can vary somewhat (from 0.04 to 0.06) depending on which words for 'earth' (*mold* or *jǫrð*) and 'meat' (*hold* or *kjöt*) we choose as the main words in Old Icelandic. This has little effect on the overall results.

[9] We thus regard changes in the original lexicon as a regular process, but the substitution of original words by loans as chance mutations that require additional correction. From this it immediately follows that glottochronological calculations become worthless in conditions where the historical and etymological situation is not sufficiently well understood for us to separate original words from loans. I would like to emphasize that no lexicostatistical survey is possible until thorough comparative work has been done. Thus, for example, the lexicostatistical calculations by Shiro Hattori, who compared the Japanese 100-word list with lists of many different languages of the world (Hattori 1959), or the calculations of Swadesh himself, who has measured the level of lexicostatistical similarity between different languages of Eurasia and America (the "Dene-Finnish" theory of Swadesh 1965), are obviously senseless.

The importance of dealing with loanwords in lexicostatistics was stressed in a recent study by Sheila Embleton (Embleton 1986), who also attempts to introduce some corrective coefficients in cases of language contacts. This is an excellent book, proving that lexicostatistics is still very much alive.

constant of 0.14 postulated by Swadesh . So it looks as if we should follow Bergsland and Vogt in assuming a slower rate of replacement in the Scandinavian languages compared to others. Now let us turn to other languages. We have at our disposal data on the development of the BL in many well-documented languages of the world, which give the following picture:

Language	t	λ_1	λ_2
Japanese	1.2	0.11	0.06[10]
Chinese	2.6	0.1	0.1[11]
English	1.3	0.14	0.1[12]
German	l.2	0.08	0.05[13]
French	1.5	0.09	0.07[14]
Spanish	1.5	0.07	0.06[15]
Rumanian	1.5	0.09	0.06[16]

This list could be enlarged: in particular it has been shown by Fodor (1961) that there has been the same low rate of development in the lexicon of the Slavonic languages. It is clear, however, that the rate of

[10] In modern Japanese, in comparison to old Japanese (eighth century A.D.) the following words have been changed: *kap-u 'eat'* > *taberu, kapigwo* 'egg' > *tamago, kasira* 'head' > *atama, isagwo* (managwo) 'sand' > *suna, wi-ru* 'sit' > *suwaru, ka-no* 'that' > *ano, nare "you"* > *anata (kimi);* and the following words have been borrowed: *kokoro* "heart" >*shinzō, kimo* "liver" > *kanzō, sisi* "meat" > *niku,* and *kapa* "skin" > *hifu, pi* "sun" >*taiyō.*

[11] In modern Chinese, in comparison with Archaic Chinese (7th century B.C.), with the total absence of loans, the following words have been replaced: *krp̂*"all" > *yiqie, duo; puk* "belly" > *duzi; ćrūʔ* "fingernail" > *jia; khwīn* "dog' > *gou, ʔəm* "drink" > *he; lək* "eat" > *chi; rōn* "egg" > *dan; ćok* "foot" > *jiao; pić* "give" > *gei; raŋ* "good" > *hao; keŋ* "neck" > *bozi; khek* "red" > *hong; wat* "say" > *shuo; kēnh* "see" > *kan; rəp* "stand" > *than; nit* "sun" > *taiyang; cə, dəʔ "*this" > *zhe; slhaj, pajʔ* "that" > *na; thəʔ* "tooth" > *ya; mōk* "tree" > *shu; nić* "two" > *liang; grāŋ* "go" > *zou; ghāj* "what" > *shemme.*

Unfortunately it is rather difficult to compile lists for intermediate stages of Chinese, since we have an accurate representation of the spoken language only in ancient texts on the one hand, and in the modern language on the other.

[12] In modern English the following words have been changed compared to Old English of the ninth century *A.D.: wamb* "belly", *micel* "big", *fuɣol* "bird", *wolcen* "cloud", *hund* "dog" (perhaps a Low German loan), ɣ*uma* (*wer*) "man", *flæsc* "meat", *heals* "neck", *weɣ* "road", *rec* "smoke", *se* "that", *þu* "you". The loans are *rinde* > *bark*, *steorfan* > *die*, *beorɣ* > *mountain*, *sinwealt* > *round*, *hyd* > *skin*.

[13] In modern German, compared to Old High German of the 8th century A.D., the following words have been changed: *bein* "bone" > *Knochen, wamba* "belly" > *Bauch, mihhil* "big" > *groß, luzzil* "small" > *klein, zagel* "tail" > *Schwanz, wīb* "woman" > *Frau.* The loans are *feizzit* "fat" > *Fett, houbit* "head" > *Kopf* and *sinwel* "round" > *rund.*

[14] In Modern French, in comparison to Vulgar Latin of the 4th or 5th Century A.D., the following words have been changed: *penna* "feather" > *plume, caput* "head" > *tête, audire* "hear" > *entendre, occidere* "kill" > *tuer, multum* "many" > *beaucoup, carnem* "meat" > *viande, arenam* "sand" > *sable, sementem* "seed" > *graine, ambulare* "go" > *aller* (possibly a loan), *natare* "swim" > *nager, parvus* "small" > *petit;* and *albus* "white" > *blanc* is a loan.

[15] In Spanish, compared to Vulgar Latin of the 4th to 5th century A.D., the following words have been replaced: *canem* "dog" > *perro* (possibly a loan), *occidere* "kill" > *matar, genuculum* "knee" > *rodilla, stare collocatum* "lie" > *estar acostado, longus* "long" > *largo, parvus* "small" > *pequen~o, ambulare* "go" > *andar, galbinus* "yellow" > *amarillo.* The loans are *pennam* "feather" > *pluma* and *albus* "white" > *blanco.*

[16] In Modern Rumanian, in comparison to Vulgar Latin of the 4th or 5th Century A.D., the following words have been changed: *ventrem* "belly" > *pîntece, grandem* "big" > *mare, avem* "bird" > *pasăre, ustulare (cremare)* "burn" > *arde, frigidus* "cold" > *rece, siccus* "dry" > *uscat, terram* "earth" > *pămînt, cordem* "heart" > *inima, buccam* "mouth" > *gura.* Loans are *collum* 'neck' > *gît, camminum* "road" > *drum, arenam* "sand" > *nisip* and *parvus* "small" > *mic*.

development of BL lists over the last one to one and a half millennia is much less than the Swadesh constant of 0.14, which can be obtained only for English and only if one treats loans as replacements[17]. We will defer for the moment our explanation of the comparatively high rate of development (0.1) of the Chinese lexicon.

Looking at the data above, it would seem easy to change Swadesh's value of 0.14 for λ to 0.06, which is the average of all λ_2 in the above cases. Taking this value of 0.06 will give dates for the Scandinavian languages which are slightly too young, and more or less exact datings for all the Romance languages, and for the remaining Germanic languages other than English. But we will get an absolutely unrealistic dating for Chinese (i.e. middle of third millenium B.C.), and the dating for Old English will be one thousand years too early. Moreover, if we use such a value of λ in formula 4 constant to date the divergences of various languages, the result is a fiasco. Thus, for the disintegration of Byelorussian and Ukrainian (which share 97% on their 100 list) we obtain:

$$t = \frac{\ln 0.97}{-2\times 0.06} = 0.25$$

That is, a separation 250 years ago, a date which obviously corresponds to nothing[18]. The same period of independent development would separate, for example, modern Persian and Tadzhik, Yiddish and German, and so on. The datings here turn out to be too young. On the other hand, if we go back to the Indo- European level and try to give dates, for example, for the period of separate development of Russian and Persian (which share 28% on a one-hundred word list), we obtain:

$$t = \frac{\ln 0.28}{-2\times 0.06} = 10.6$$

— that is, the beginning of the ninth millennium B.C. But the more or less established view is that the disintegration of common Indo-European took place in the fourth millennium B.C. Other pairs of Indo-European languages will also yield datings that are far too early.

Now we can understand why Swadesh chose a value of 0.14 for λ. With a normal exponential correlation between time and percentage retention this value avoids giving us unreasonably old dates at great time depths, while at the same time giving slightly too recent dates for the whole period with externally verifiable datings. We have already seen that the value of 0.14 is in reality found only in English, and even there some caveats regarding borrowings are needed. It follows that it cannot have the status of an empirically established value. On the other hand, the empirically observed value of λ allows us to give reasonable datings for events which happened one millenium or one and a half millenia ago, but is absolutely unsuitable for dating earlier or later splits. We need to understand the nature of this contradiction and to find some kind of solution.

Let us once more turn to the case of the Romance languages. We estimated above the rate of

[17] In the remaining languages, even if we count the loaned items, the rate is substantially less than 0.14. The somewhat 'accelerated' development of English should be probably explained by active language interference during the formation of Modern English, with Scandinavian and Romance components playing a part.

[18] For the separation of Belorussian and Ukrainian we obtain a date of the mid nineteenth century using the Swadesh formula; for the separation of Icelandic and Faroese, as mentioned above, the eighteenth century. For the disintegration of proto-Slavic, of proto-Germanic, and proto Romance (i.e. Vulgar Latin) we obtain datings around the 11th or 12th centuries A.D., and for the disintegration of Common Indo-European, the middle or end of the third millennium B.C.

replacement on the basis of a comparison of the BL lists of the Romance languages with the BL of Vulgar Latin, the proto-language of all modern Romance languages which disintegrated around the fourth to sixth centuries A.D. Let us now compare the BL of classical Latin, which dates from the fourth to second centuries B.C. and has a well-known BL[19], with a view to what changes took place in the French, Spanish and Rumanian BLs in comparison with classical Latin.

The following words have been changed in French: *omnis > tout* 'all', *magnus > grand* 'big', *edere > manger* 'eat', *pinguedo (adeps) > graisse* 'fat', *penna > plume* 'feather', *ignis > feu* 'fire', *caput > tête* 'head', *audire > entendre* 'hear', *occidere > tuer* 'kill', *cubare > être couché* 'lie', *jecur > foie* 'liver', *vir > homme* 'man', *multus > beaucoup* 'many', *caro > viande* 'meat', *os > bouche* 'rot', *arena > sable* 'sand', *semen > graine* 'seed', *cutis > peau* 'skin', *parvus > petit* 'small', *nare (natare) > nager* 'swim', *ire > aller* 'go', flavus > *jaune* 'yellow'. The following are loans: *via > chemin* 'road', *lapis > pierre* 'stone' and *albus > blanc* 'white'. We thus have a 77% retention of lexicon; letting t ≈ *2.3* we get a value of λ*t = 0. 11,* whereas the rate between Vulgar Latin and Modern French is λ = 0.07 — see above.

In Spanish the following words have changed: *omnis > todo* 'all', *magnus > grande* 'big', *urere > quemar* 'burn', *canis > perro* 'dog', *pinguedo (adeps) > grasa* 'fat', *ignis > fuego* 'fire', *occidere > matar* 'kill', *genu > rodilla* 'knee', *cubare > ester acostado* 'lie', *jecur > higado* 'liver', *longus > largo* 'long', *vir > hombre* 'man', *os > boca* 'mouth', *via > camino* 'road', *cutis > piel* 'skin', *parvus > pequeño* 'small', *ire > andar* 'go', *flavus (fulvus) > amarillo* 'yellow'. The following are loans: *penna > pluma* 'feather', *lapis > piedra* 'stone', and *albus > blanco* 'white'. We thus have a retention rate of 80%; letting *t = 2.3* we get a value of λ*t = 0.1,* whereas the rate between Vulgar Latin and Modern Spanish is λ = 0.06 — see above.

In Rumanian the following words have changed: *omnis > tout* 'all', venter > *pîtece* 'belly', *magnus > mare* 'big', avis > *pasăre* 'bird', *urere > arde* 'burn', *frigidus > rece* 'cold', *siccus > uscat* 'dry', *terra > pămînt* 'earth', *edere > mînca* 'eat', *pinguedo (adeps) > grasime* 'fat', *ignis > foc* 'fire', *cor > inima* 'heart', *cubare > sta culcat* 'lie', *jecur > ficat* 'liver', *os > gură* 'mouth', *vir > om (barbat)* 'man', *cutis > piele* 'skin', *ire > umbla 'go', flavus (fulvus) > galben* 'yellow'. The loans are: *collum > gît* 'neck', *via > drum* 'road', *arena > nisip* 'sand', *parvus > mic* 'small', and *lapis > piatra* 'stone'. This is the same retention rate of 80% as Spanish, which gives us the same rate over 2.3 thousand years of λ*t* = 0.1, again compared to the rate of λ = 0.06 between Vulgar Latin and Rumanian as shown above.

We see, then, a clear case of the rate of change increasing (i.e. an acceleration) with an increased rate of lexical loss in BL as the separation time λ*t* increases. This fact explains the extremely low rate of lexical loss since the separation of Belorussian and Ukrainian, and also between Persian and Tadzhik: we get 97% of correspondences between each of these pairs, both of which separated about 600 years ago, and thus λ should be set at about 0.03 for these cases[20]. At the same time, this is the explanation of the comparatively high figure of 0.1 for the rate of loss within the Chinese lexicon: the time elapsed between Archaic and Modern Chinese is about the same as that between Classical Latin and the modern Romance languages.

As mentioned above, the mathematical apparatus of glottochronology was borrowed from the theory of radioactive decay. But the most important difference between words and neutrons is perhaps the fact that the former, by contrast with the latter, can become 'older'. In fact, the probability of a neutron

[19] Vulgar Latin is dated differently by different scholars (cf. Muller 1929, Gurycheva 1959). But it is obvious that it had a unitary nature until the fifth century A.D. despite the presence of some dialectal differences. It began to separate into dialects between the fifth and eight centuries A.D., and the period of Romance languages dates back to the eight century.

[20] The separation of Belorussian and Ukrainian can be dated by historical records to the thirteenth or fourteenth centuries, if one correlates this with the separation of Belorussia after its conquest by Lithuania. For the separation of Persian and Tadzhik the probable dating is the end of the fourteenth or the beginning of the fifteenth century, when Tadzhikistan and the eastern part of Iran separated as a result of the Mongolian invasion.

remaining intact at a given time is always $e^{-\lambda t}$, regardless of its history. On the other hand, the probability of a word (including those in BL) remaining correlates with how long this word has already 'lived'.

We can thus explain, for example, the replacement in all modern Indo-European languages of the proto Indo-European word *g^{w}(e)ru-* 'heavy': this word is represented in Vedic *guru-*, Ancient Greek βαρύς, Lat. *gravis,* Gothic *kaurus,* but is lost in the majority of modern languages: with the meaning 'heavy' this stem is maintained only in Modern Greek and in some modern Indian languages, while in most Indian, Germanic, Romance, Slavonic and other Indo-European languages it has either changed its meaning or disappeared. Comparativists are familiar with many examples of this kind -namely the wide distribution of some words or roots in ancient languages and their almost total absence from the modern languages of the same family — which seem to be connected to the 'lifetime' of given words.

The suggestion that words can 'age' automatically leads us to reject the third postulate of glottochronology — that is, of the idea that the rate of retention in BL is constant — and to adopt the hypothesis that there is a correlation between λ and time *t*. And in fact if we assume that the greater the value of time t then the higher the probability that an original word from BL will disappear, then instead of (1) we adopt as our main formula in glottochronology

$$(5)\ N(t) = N_o e^{-\lambda t2}$$

This is the formula for regular acceleration of speed[21].

Now, the time *t* can be established by the following formula:

$$(6)\ t = \sqrt{\frac{\ln N(t)}{-\lambda \times N_o}}$$

and the time of separation between n languages by the formula

$$(7)\ t = \sqrt{\frac{\ln Nn(t)}{-n\lambda N_o}}$$

If we insert the data given above for the various languages in this formula we obtain the coefficient of acceleration λ:

Language	T	λ
Japanese	1.2	0.05
English	1.3	0.08
Chinese	2.6	0.04
German	1.2	0.04
French	1.5	0.05
Spanish	1	0.06
Riksmål	1	0.05

[21] The possibility of such a correlation was already envisaged, though only theoretically, in Arapov & Herz 1974.

With the exception of English, which reveals a somewhat higher value, the value for λ is stable and varies only slightly between 0.04 and 0.06. In addition, when we date language separation over the last thousand to one and a half thousand years the above formulae on the whole give good results. Thus for Belorussian and Ukrainian as well as for Persian and Tadzhik we have:

$$t = \sqrt{\frac{\ln 0.97}{-2\times 0.05}} = 0.55$$

(i.e. the end of the fourteenth century A.D.)

For the date of separation of Icelandic and Faroese (94% correspondence rate) we get about the twelfth century instead of the eighteenth century according to the Swadesh formula, and so on.

However, one can show that the use of formula (5), when applied to separations of greater time depth, begins to yield dates that are too recent. Thus for Russian and Persian the application of (7) with a value of 0.05 for λ, the time of separation would be

$$t = \sqrt{\frac{\ln 0.28}{-2\times 0.05}} = 3.6$$

i.e. around the middle of the second millenium B.C. Even with the help of the Swadesh formula, with 28% shared lexicon we will obtain the end of the third millenium B.C. whereas the real dating should rather be the fourth millenium B.C. (see above).

Moreover, one can show that if we adopt formula (5) the period for the complete replacement of the BL should be less than 10,000 years, and for any two separated languages all correspondences on the one hundred word list should have disappeared after seven millennia. There is no doubt that, if we adopt the formula

(5) $N(t) = N_o\, e^{-\lambda t2}$

the dating will be somewhat improved for recent periods, and will agree better with real instances of lexical development. But for more ancient periods it will give results that are worse than those of Swadesh. How do we resolve this contradiction?

In order to advance further, we need to discuss the fourth postulate of glottochronology: 'all words of BL have the same chances of being retained over a given *t*' (above). It was this statement that aroused the strongest objections both among supporters of glottochronology and its detractors. In fact, any linguist who works with glottchronology knows that not all words in BL are of equal stability. The words 'small' or 'skin' have on the whole a better chance of disappearing from the list than such words as 'I', 'you' or 'ear'.

Therefore it has often been suggested that the coefficient of the retention rate for the whole list should in general be derived from the individual coefficients of retention for each word, i.e. the probability of it being retained over a given period of time t (see, e.g., Van der Merwe 1966). An attempt to determine such individual coefficients empirically felt (Austronesian and Indo-European languages) has been made by Dyen and James (1967), but their mathematical apparatus was too baroque and unproductive. Moreover, if the suggestion that different words have different retention rates is true, it is also quite probable that these individual coefficients of retention could vary according to the cultural and linguistic environment. For example, such words as 'cloud' and 'tail' are very stable in the Turkic languages but unstable in Germanic;

the word 'belly' is very stable in Romance but unstable in Slavic and so forth. Although some words from the BL do indeed reveal a high level of stability ('I', 'you', 'sun', 'eye', 'ear' etc.), it is clear that a formula based on individual coefficients cannot solve this problem, because it will not be universally applicable.

Let us imagine now an ideal BL list with some average rate of divergence λ, where all words are ranked according to the probability of their disappearance over a given period of time Δ*t*, with the first word in the list having a probability of being retained close or equal to zero, while the last word has a probability approaching 1. In this case, words should disappear in turn, beginning with the least stable and going on to the more stable. Accordingly, at time t_n, the rate of loss for those items remaining on the list should become slower than at time t_{n-1} . The rate of loss of this list will, it follows, be variable, depending at any moment on the proportion of retained (and therefore of lost) words:

$$(8)\ \lambda t_n = \lambda_o N(t_n)$$

If we put (8) in formula (5) we obtain the formula:

$$(9)\ N(t) = N_o\, e^{-\lambda N(t)t2}$$

Time (t) can be calculated according to the formula:

$$(10)\ t = \sqrt{\frac{\ln N(t)}{-\lambda \times N(t) \times N_o}}$$

Note that in the case where we have n languages, the formula giving the date of their disintegration can be recast as:

$$(11)\ N_n(t) = N_o\, e^{-n\lambda N(t)t2}$$

where N(t) is the value, in reality not attested, but obtainable more or less precisely by the formula:

$$(12)\ N(t) = \sqrt[n]{Nn(t)}$$

From (11) and (12) we get a means of calculating the time of divergence for a proportion of corresponding words in the BL of n languages:

$$(13)^{22}\quad t = \sqrt{\frac{\ln \left(\frac{Nn(t)}{N_o}\right)}{-n\lambda \sqrt[n]{Nn(t)}}}$$

[22] For practical purposes (in the most frequent case, when n = 2, $N_0 = 1$ (100 words) with λ = 0.05 and with a more traditional rendering of N(t) as c) the formula may be simplified as:

$$t = \sqrt{\frac{\ln c}{-0.1\sqrt{c}}}$$

Formula (9) represents the 'contradictory' character of the process of lexical loss in BL: the square of *t* reflects the acceleration of loss caused by the 'aging' of words, while the coefficient *N(t)* in the exponent reflects the opposite to this — the deceleration of the rate of loss when less stable words disappear from the remaining BL list and the more stable items are retained[23]. We may note that for small values of *t* (and accordingly for large *N(t)),* the datings according to formulae (9) and (5) will be similar. In contrast, as *t* grows (and *N(t)* diminishes) the datings become significantly earlier. If we assume that λ is 0.05 on average (see above), we get the following table of selected datings, where *N(t)* is the proportion of words from BL remaining in one language, $N_2(t)$ is the proportion of words corresponding between two languages, *t is* the time of development and therefore of divergence in millennia:

N(t)	$N_2(t)$	t	t (According to M. Swadesh)
0.99	0.99	0.3	0.03
0.97	0.94	0.8	0.2
0.95	0.9	1	0.35
0.9	0.81	1.5	0.7
0.85	0.72	2	1.1
0.8	0.64	2.4	1.5
0.75	0.56	2.8	1.9
0.7	0.49	3.2	2.4
0.65	0.42	3.7	2.9
0.6	0.36	4.1	3.4
0.55	0.3	4.7	4
0.5	0.25	0.53	4.6
0.45	0.2	6	5.3
0.4	0.16	6.8	6.1
0.35	0.12	7.8	7
0.3	0.09	9	8
0.25	0.06	10.7	9.3
0.2	0. 04	12.7	10.7
0.15	0.02	16.6	13
0.1	0.01	21.5	15.3

It appears that the datings obtained by (13) are much more reliable than those of classical glottochronology. For example, we date the disintegration of Belorussian and Ukrainian (97% correspondences) to be the fourteenth century A.D.; for different pairs of Germanic, Romance, Slavic or Turkic languages we get datings in the first millennium; for the disintegration of the Balto-Slavonic languages we get a dating around the end of the second millennium B.C. Of course, one should not overemphasize the precision of glottochronological data, since there may be various types of statistical fluctuation and perturbation[24].

[23] A thorough criticism of the mathematical apparatus of glottochronology was undertaken by Chrétien (1952). However all his objections were effectively countered by Dyen. We will therefore not discuss Chrétien's objections here, but simply note that they have no connection with the problem of the variability of retention rates.

[24] Language contact is an obvious disturbing factor. This can produce secondary linguistic convergence of two already diverged languages. In this case often the fifth postulate of glottochronologv is violated: in languages which are in active contact there is a tendency towards retention and/or loss of the same words from the BL. In such cases we cannot always speak of borrowing. Sometimes such contacts can cause an increase in the percentage of retentions by as much as 5 or 6 percent. We can observe this situation for Byelorussian and West Slavic languages, for German and

However, it seems to be an important adjunct in genetic classification and in the evaluation of the closeness of related languages.

Everything said above relates to Swadesh's 'standard' one-hundred word BL. The existence of differences in individual retention rates suggests that, in principle, it is possible to compile lists which would differ from each other not only in the general coefficient λ, but also in their formulae for the correlation of rate of loss and time. It is not impossible that by varying the set of words in the list one could compile lists which satisfy different correlation formulae. This work, however, would be computationally complex[25].

Classical lexicostatistics, even with these improved dating methods, is still plagued by major shortcomings. There are well-known problems connected with the choice of the main word where two or more synonyms exist — for example, what word should be chosen in Italian for 'head' — *capo* or *testa ?* — or 'sand' — *sabbia* or *rena ?* — and so on. Apart from that, the lexicostatistical procedure lacks statistics in the strict sense: in comparing the BLs of two languages, we get but a single result, while to increase the reliability of results one would like to have a series of outcomes, from which it would be possible to calculate the mathematical confidence and the limits of possible fluctuation.

I assume that both these shortcomings could he eliminated with the help of a method which I call 'etymological statistics', or 'root glottochronology'. Let us formulate its main postulates.

1. In every language there are some roots that are original, i.e. not borrowed during the period of separate existence of this language. According to preliminary estimates, there are not much more than two or three thousand roots of this type in any modern language.

2. These roots have different frequencies of occurrence, in other words they have different probabilities of being found in a chosen text.

3. The frequency of occurrence (as just defined) of a given root in some language at a fixed period of time t is stable, and does not depend (or hardly depends) on the type of text.

4. All roots can 'age' — their frequency of occurrence gradually approaches zero, after which the root is considered to have disappeared from the language. At the same time, however, the rate of loss of different roots is not identical: roots, like words, may be divided into stable and less stable.

5. The loss of roots from a language proceeds at a steady rate — that is, from some set of roots,

Scandinavian, etc. There are some ways of taking account of this effect while evaluating the degree of closeness, but all of these have a restricted application and they deserve separate consideration. An attempt of a general discussion of this problem was suggested by Hattori Shiro (1954) who suggested the constant 2 in the formula $N_2(t)=e^{-2\lambda t}$ be replaced by a variable. This variable depends upon the degree of closeness between the languages. However it remains obscure how this coefficient is to be established.

[25] In an experiment we combined words from the 100 word list as well as from the 200 word list of M. Swadesh and tried to compile a 55 word list which was to fulfil the classical "radioactive" equation $N(t) = Ne^{-\lambda t}$ where $\lambda = 0.1$ (i.e. "where the value of the coefficient of persistence was 0.9 for 1,000 years). This list contained the following words: *bark, belly, big, black, blood, bone, to burn, to die, dog, dry, ear, to eat, egg, eye fire, foot, full, hair, head, I, knee, leaf, liver, long, many, meat, moon, near, night, nose, round, short, snake, star, stone, swim, tail, this, thin, thou, tongue, tree, two, water, we, what, white, woman, worm, year, yellow, who, neck, new, mouth*. We tried to compile this list in such a way that it satisfied the procedure of 'classical' glottochronology (in particular, borrowings were counted as replacements). This list, which has been tested on relatively large linguistic material, allows us to build classificatory schemes which are reliable on the whole, and to find datings which match the datings according to the standard list of Swadesh (although with sometimes quite considerable deviations because of the smaller number of words); for short historical time spans though, datings still are definitely too 'young'.

The list of 35 most stable words with $\lambda = 0.07$ or 0.08 has been compiled by Yakhontov. This list is very useful for verification of genetic relationship between languages. The percent of cognates among these 35 words should be higher than the percent between the remaining words of the list. However, this short list is not quite suitable for dating or classification.

characterized by a fixed frequency, over a given period Δt, a fixed number of roots will be lost.

For this theory, the key postulate is 3, and a priori it raises most doubts. Indeed, the nature of the lexicon and the distribution of word frequency in texts of various genres vary considerably, except for such extremely frequent words as pronouns and grammatical morphemes. But one etymological root usually serves as a source for many derived words with different meanings. Since the frequency of a root correlates strongly with its productivity, the most frequent roots are usually found in stems belonging to very different semantic domains. Thus the higher the productivity and frequency of a root the better the chance of finding it in texts of any genre and subject matter. In any case, the empirical data clearly demonstrate the genre-independent and neutral character of the set of roots found in any text.

The fifth postulate of 'root glottochronology' is equivalent to the third postulate of standard lexicostatistics (see above), and requires empirical confirmation (see below).

Let us take some samples — for example, one hundred non-loaned roots from a language A — and let us supply each of these roots with their etymological correspondents in related languages. (I should note that this procedure necessarily presupposes prior etymological analysis). It is obvious that in those languages that are most closely related to language A, one can find the highest number of correspondences, while with increasing genetic distance the number of such correspondences will diminish. It is natural to expect that Russian, for example, will have more common roots with Slavic languages than with Baltic, and with the latter — more than with all other Indo-European languages, etc. This procedure would allow us to create classifications and relative chronologies of divergence within a language family. The analysis of several such samples should in principle give comparable results.

However, due to postulates (2) and (4), the absolute figures of etymological correspondences among roots of language A and related languages will vary considerably in random samples. In order to make these figures stable, one should select samples characterized by one and the same distribution of root frequencies.

One could compile dictionaries of root frequency and take one's samples from them. This would mean extending to the study of roots the work carried out for words by Arapov and Herz. However, such work is difficult to accomplish.

Here postulate (3) comes into play, according to which each root in the language has a given probability of being found in any text. Accordingly, any text should exhibit the same or similar distribution of frequencies of the roots represented in it.

If this is so, one would expect that in samples of genuine roots in different texts from language A there would be the same or a similar number of etymological correspondences with each of the related languages.

Let us take some English text — for example, the text of this article. We will choose from it non-loaned lexical roots; all prefixes, suffixes and proper names will be excluded, and each root will be counted only once. For each morpheme of this type we shall look for etymological correspondences in German, Russian, Lithuanian, and French. We will not count cases where etymological correspondences are present in these languages, but only as loans.

	English	German	Russian	Lithuanian	French	IE
1.	the, this,that	+ (der)	+ (тот)	+ (tà-s)	+ (te-l)	*to-
2.	last	+ (letzt)	+ (лень)	+ (léid-)	+ (las)	*lē(i)-
3.	two	+ (zwei)	+ (два)	+ (dù)	+ (deux)	*du̯ō-

4. have	+ (haben)	-	-	+ (re-cev-)	*kap-
5. wit-ness	+ (wissen)	+ (вед-)	+ (vīd-)	+ (voir)	*weid-
6. a, once, any	+ (ein)	+ (од-ин)	+ (víenas)	+ (un)	*oino-
7. in	+ (in)	+ (в)	+ (ĩ)	+ (en)	*en-
8. of	+ (ab)	+ (по)	+ (pa-, po)	+ (po-ndre)	*apo/*po
9. work	+ (wirken)	-	-	-	*werǵ-
10. with	+ (wider)	+ (второй)	-	-	*wi-
11. wide	+ (weit)	-	-	-	(*weit-)
12. laid	+ (legen)	+ (лежать)	-	+ (lit)	*legh-
13. for, from	+ (für)	+ (про, при)	+ (pro)	+ (pour)	*per, *pro-
14. or , other, be-yond, and	+ (ander-,jener)	+ (он)	+ (ana-s)	-	*eno-, *no-
15. step	+ (Stapfe)	-	-	-	*steb-
16. was	+ (war)	-	-	-	*wes-
17. by	+ (bei)	+ (o[b])	-	+ (oub-lier)	*(o)bhi
18. his, hinder	+ (hier, hinter)	+ (з-десь)	+ (šìs)	+ (ce)	*ḱe-, *ḱi-
19. to-wards	+ (werden)	+ (вертеть)	+ (veȓsti)	+ (vers)	*wert-
20. which, who	+ (wer, welcher)	+ (к-то, ч-то)	+ (kà-s)	+ (que)	*k^wo-
21. begin	+ (beginnen)	-	-	-	(*ghen-?)
22. to	+ (zu)	+ (до)	+ (-da)	+ (de)	*do-/*de-
23. at	-	-	-	+ (à)	*ad
24. old	+ (alt)	-	-	+ (haut)	*al-
25. world	+ (Welt)	-	-	+ (vertu)	*wīro-
26. new	+ (neu)	+ (новый)	+ (naujas)	+ (neuf)	*newo-

27. if , it	+ (ob, es)	+ (э-тот)	+ (jì-s)	+ (ce < ecce)	*e-, *i-
28. we	+ (wir)	-	+ (vè-du)	-	*we-
29. as, also, all	+ (als, all)	-	+ (al-víenas)	-	*al-
30. sixth	+ (sechs)	+ (шесть)	+ (šešì)	+ (six)	*s(w)eḱs-
31. through	+ (durch)	-	-	+ (très, tra-)	*ter-
32. out, b-ut, a-b-out	+ (aus)	+ (вы-)	+ (už-)	+ (j-usque)	*ud-
33. is	+ (ist)	+ (есть)	+ (es-)	+ (es-t)	*es-
34. are	-	-	+ (yrà)	-	*er-/*or-
35. arise	+ (reisen)	+ (рой)	+ (rý-tas)	+ (ruisseau)	*(o)rei-
36. so, such	+ (so, sich)	+ (свой, себя)	+ (sãvo)	+ (se)	*s(w)e
37. can, know	+ (kann)	+ (знать)	+ (žinó-ti)	+ (connaître)	*ǵnō -
38. must	+ (mu-)	-	-	+ (co-mme)	*med-
39. now	+ (nun)	+ (ныне)	+ (nù)	-	*nŭ-
40. answer	+ (schwören)	+ (свара)	-	-	*swer-
41. be	+ (bi-n)	+ (быть)	+ (bū-ti)	+ (fu-t)	*bheu-
42. same, some	+ (zu-sammen)	+ (сам)	+ (san-)	+ (sim-ple)	*sem-
43. rather	+ (retten)	-	+ (krẽs-ti)	-	*kret-
44. would, well	+ (wollen)	+ (воля)	+ (valià)	+ (vouloir)	*wel-
45. not	+ (nicht)	+ (не)	+ (nè)	+ (non)	*ne
46. our, us	+ (uns)	+ (нас)	-	+ (nous)	*no-(s)
47. kind	+ (Kind)	-	-	+ (naître)	*ǵenə-
48. least	-	-	-	-	(*leis-)
49. should	+ (sollen)	-	+ (skelḗ-ti)	-	*skel-
50. earlier	+	-	-	-	*ai̯er-

	(eher, erst)				
51. much	-	-	-	+ (maire, mais)	*maǵ- (*meǵ)
52. more, most	+ (mehr, meist)	-	-	-	*mē-
53. on	+ (an)	+ (на)	+ (nuõ)	-	*an-a, *an-ō
54. ever, every	+ (ewig)	-	-	+ (âge)	*aiw-
55. ten	+ (zehn)	+ (десять)	+ (dẽšimt)	+ (dix)	*deḱm-
56. thousand	+ (Tausend)	+ (тысяча)	+ (tū́kstantis)	-	*tūs-
57. year	+ (Jahr)	+ (яровой)	-	-	*i̯ōr-
58. time	+ (Zeit)	-	-	-	*dā(i̯)-
59. depth	+ (tief)	+ (дно)	+ (dubùs)	-	*dheub-
60. in-deed, do	+ (tu-n)	+ (де-ть)	+ (dė́-ti)	+ (fai-re)	*dhē-
61. small	+ (schmal)	+ (малый)	-	+ (mal, mauvais)	*(s)mă̆l-
62. greater	+ (gro-)	+ (груда)	+ (grús-ti)	-	*ghreud-
63. like-wise	+ (gleich)	-	+ (lýgus)	-	*līg-
64. deal	+ (Teil)	+ (делить)	+ (daily´-ti)	-	*dhoi-l-
65. call	+ (klagen)	+ (голос)	-	-	*gol(-s-)
66. tree	+ (treu, Teer)	+ (дерево)	+ (dervà)	-	*derw-
67. wood	-	-	-	-	*widhu-
68. few	-	-	-	+ (pauvre)	*pau-
69. find	+ (finden)	+ (путь)	-	+ (pont)	*penth-
70. root	+ (Wurzel)	-	-	+ (racine)	*wərad-
71. side, since	+ (Seite, seit)	-	-	+ (po-nd-re)	*sēi̯-
72. back	-	-	-	-	(*bhag-)
73. may	+ (mag)	+ (мочь)	+ (mė́g-ti)	-	*megh-

74. stem	+ (Stamm)	-	+ (stem̃b-ti)	-	*ste(m)bh-
75. set, settle	+ (sitzen)	+ (сидеть)	+ (sēdḗ-ti)	+ (as-soire)	*sed-
76. even	+ (eben)	-	-	-	(*ebh-?)
77. still	+ (still)	+ (стол, столб)	+ (stãlas,stul̃bas)	+ (lieu)	*stel-
78. East	+ (Osten)	-	+ (aušrà)	-	*awes-
79. West	+ (Westen)	-	-	+ (vêpres)	*wes-
80. hand	+ (Hand)	-	-	-	(*kon-/*ḱ-, -m-)
81. ninth	+ (neun)	+ (девять)	+ (devynì)	+ (neuf)	*new-m
82. ask	+ (heischen)	+ (искать)	+ (ieškó-ti)	-	*ais-
83. over, often	+ (über)	+ (высокий)	-	+ (sous)	*upo
84. give	+ (geben)	-	+ (gabénti)	+ (avoir)	*ghabh-
85. young	+ (jung)	+ (юный)	+ (jáunas)	+ (jeune)	*i̯ewən-
86. belong	+ (lang,gelangen)	+ (долгий)	+ (ìlgas)	+ (long)	*delə(n)gh-
87. many	+ (manch)	+ (много)	+? (minià)	-	*menegh-
88. higher	+ (hoch)	+ (куча)	+ (kaũkaras)	-	*keuk-
89. show	+ (schauen)	+ (чуять)	-	-	*(s)kew-
90. fully	+ (voll)	+ (полный)	+ (pìlnas)	+ (plein)	*pelə-
91. I	+ (ich)	+ (я)	+ (àš)	+ (je)	*eǵ(h)om
92. think	+ (denken)	-	-	-	*tong-
93. word	+ (Wort)	+ (врать)	+ (var̃das)	+ (verve)	*wer-
94. half	+ (halb)	-	-	-	(*kalp-)
95. a-go	+ (gehen)	-	-	-	*ǵhē-
96. share	+ (scheren)	-	+ (skìr-ti)	-	*(s)ker-

97. five	+ (fünf)	+ (пять)	+ (penkì)	+ (cinq)	*penk^{w}e
98. start	+ (stürzen)	+ (страдать)	+ (starìnti)	+ (étrenne)	*ster-
99. talk	+ (Zahl)	-	-	+ (deuil)	*del-
100. say	+ (sagen)	-	-	-	*sagh-

Thus in this sample of text among 100 non-loaned roots[26] we find (if we score doubtful etymologies at 0.5%) 94 roots shared with German, 58.5 with Lithuanian, 60 with Russian, and 58 with French.

It turns out that these figures are quite stable. Let us compare the results of the calculations done with some Russian texts (with the text number in brackets — see page below):

	Polish	Lithuanian	German	French
(4)	98	77	55	52
(1)	96	76	58	50
(5)	97	70	51.5	50
(11)	93	72	55	51.5

We see that the figures in each case cluster around some statistical mean: for Polish it is 95 ± 3%, for Lithuanian 74 ± 3%, for German 54 ± 3%, and for French 51 ± 1%.

We shall now give some more results for different pairs of languages. By 'text language' we mean the language whose text is being analyzed, and by 'dictionary language' the language whose dictionary is used for the comparison.

Text Language	Dictionary Language	Proportion of Correspondences	Text
French	Russian	0.50	7
German	Russian	0.55	6
German	Lithuanian	0.57	6
Latin	Russian	0.50	8
Latin	German	0.55	8
Latin	Lithuanian	0.53	8
Ancient Greek	Vedic	0.69	9
Ancient Greek	Latin	0.67	9
Ancient Greek	Russian	0.52	9
Ancient Greek	German	0.54	9
Ancient Greek	Lithuanian	0.53	9

[26] From 100 selected morphemes, *begin, back, even, hand* and *half* do not have IE etymologies. Nevertheless it is not proven that these morphemes have been borrowed. Thus we have retained them in the list. In any case, there are too few of them to have any significant influence on the results of our analysis. On the other hand, we excluded from the list many stems whose status as borrowings is well-proven. An average English text contains in fact many more borrowings than inherited roots which may be actually the reason for a slight increase of the cognacy rates, compared with respective figures for Russian, German and other texts.

Text Language	Dictionary Language	Proportion of Correspondences	Text
Latin	Vedic	0.71	8
Latin	Ancient Greek	0.72	8
Russian	Lithuanian	0.76	3
Russian	Lithuanian	0.74	2
Vedic	Ancient Greek	0.77	10
Vedic	Latin	0.64	10
Vedic	Russian	0.54	10
Vedic	Lithuanian	0.53	10
Vedic	German	0.57	10

Besides the relatively high stability of correspondence levels, we observe an interesting phenomenon. While calculating from 'text language' to 'dictionary language' and vice versa (i.e. when the roles of 'text language' and 'dictionary language' are exchanged) the figures for languages of the same period are similar. That the figures should increase, when comparing modern 'text languages' with ancient 'dictionary languages' is only to be expected. But the calculations from ancient 'text languages' to modern 'dictionary languages' reveal practically the same figures as when we compare two modern languages. One can thus formulate a very important rule: the age of the text does not influence the result of the statistical analysis.

With closer consideration this result becomes quite understandable. The conditions required by the method in each case let us measure not the distance from one language to another, but the distance from the proto-language to the dictionary language[27], and therefore we obtain only the figures characterizing this distance.

Let us try one more tack. We shall take as a sample text the one-hundred word Swadesh list and will apply the same procedure to it. That is, we will write out all non-loaned roots, ignoring borrowings and repetitions. Let us analyse this material with the help of the same languages as above, comparing German, Russian, Lithuanian and French with English.

	English	German	Russian	Lithuanian	French	IE
1.	all	+ (als, all)	-	+ (al-v́ienas)	-	*al-
2.	ashes	+ (Asche)	-	-	+ (ardeur)	*as-
3.	belly	+ (Balg)	-	-	-	*bhelǵh-
4.	bird	-	-	-	-	(*bhreu-)?
5.	bite	+ (bei-en)	-	-	+ (fendre)	*bheid-
6.	black	+ (blecken)	-	-	+ (foudre)	*bhelg-
7.	blood	+ (Blut)	-	-	+ (fleur)	*bhelə-

[27] Translator's Note: Since we are in fact measuring how many roots reflected in the sample are retained in the dictionary language (*NE & IP*).

8.	bone	+ (Bein)	-	-	-	(*bhoin-)
9.	breast	+ (Brust)	+ (брюхо)	-	-	(*bhreus-)
10.	burn	+ (brennen)	+ (бруить)	+ (briáutis)	-	*bhreu-
11.	cloud	-	-	-	-	*gleut-
12.	cold	+ (kalt)	+? (холод)	+? (šáltas)	+ (gel)	*gel-
13.	come	+ (kommen)	-	+? (gim̃ti)	+ (venir)	*g^{w}em-
14.	die	+ (tot)	-	-	-	*dheu-
15.	drink	+ (trinken)	-	+ (drė́ž-ti)	-	*dhreǵ-
16.	dry	+ (trocken)	-	-	-	*dhreugh-
17.	ear	+ (Ohr)	+ (ухо)	+ (ausìs)	+ (oreille)	*ous-
18.	earth	+ (Erde)	-	-	-	*er(t)-
19.	eat	+ (essen)	+ (есть)	+ (ē´sti)	-	*ed-
20.	eye	+ (Auge)	+ (око)	+ (akì-s)	+ (oeuil)	*ok^{w}-
21.	fat	+ (feist)	+ (питать)	+ (py´dyti)	-	*poi-
22.	feather	+ (Feder)	+? (перо)	+ (spar̃nas)	+ (penne)	*pet-/ pter-[28]
23.	fire	+ (Feuer)	-	-	-	*peHwōr
24.	fish	+ (Fisch)	+? (пескарь)	-	+ (poisson)	*peisk-
25.	fly	+ (fliegen)	+ (плыть)	+ (pláuti)	+ (pluie)	*pleu-
26.	foot	+ (Fu-)	+ (под)	+ (pãdas)	+ (pied)	*ped-
27.	full	+ (voll)	+ (полный)	+ (pìlnas)	+ (plein)	*pelə(n)-
28	give	+ (geben)	-	+ (gabénti)	+ (avoir)	*ghabh-
29.	go	+	-	-	-	*ǵhē-

[28] As we have Greek πτεϱόν, it seems reasonable that the IE word for "feather" was originally connected with IE **pet-* "to fly". Due to development from **pter-* to **per-* in different languages including Balto-Slavic, there were opportunities of contamination with the IE root **per-* "to move", Russian *parit′* and German *fahren* etc.

		(gehen)				
30.	good	+ (gut)	+ (годный)	+ (guõdas)	-	*ghadh-
31.	green	+ (grün)	-	-	-	*ghrē-
32.	hair	+ (Haar)	+ (шерсть)	+ (šerỹs)	-	*ḱer(s)-
33.	hand	+ (Hand)	-	-	-	(*kon-/*ḱ-, -m-)
34.	head	+ (Haupt)	-	-	+	*kap-
35.	hear	+ (hören)	+ (чуять)	-	-	*(s)kew-
36.	heart	+ (Herz)	+ (сердце)	+ (širdìs)	+ (coeur)	*ḱerd-
37.	horn	+ (Horn)	+? (корова)	+? (kárve)	+ (corne)	*ḱern-
38.	I	+ (ich)	+ (я)	+ (àš)	+ (je)	*eǵhom
39.	kill	+ (quälen)	+ (жаль)	+ (gélti)	-	*g^{w}el-
40.	knee	+ (Knie)	+ (звено)	-	+ (genoux)	*ǵenu-
41.	know	+ (kennen)	+ (знать)	+ (žinóti)	+ (connaître)	*ǵnō -
42.	leaf	+ (Laub)	+ (луб)	+ (lubà)	+ (livre)	*leubh-
43.	lie	+ (liegen)	+ (лежать)	-	+ (lit)	*legh-
44.	liver	+ (Leber)	-	-	-	*lepro-
45.	long	+ (lang)	+ (долгий)	+ (ìlgas)	+ (long)	*delə(n)gh-
46.	louse	+ (Laus)	+? (вошь)	+? (v̀ievisa)	-	*wes- / *li̯us-
47.	man	+ (Mann)	+ (муж)	-	-	*mon-
48.	many	+ (manch)	+ (много)	+? (minià)	-	*menegh-
49.	meat	+ (Mast)	-	-	-	*mad-
50.	moon	+ (Mond)	+ (месяц)	+ (mḗnuo)	+ (mois)	*mēnes-
51.	mouth	+ (Mund)	-	-	+ (menton)	*men-

52.	nail	+	+	+	+	*(o)nogh^w-
		(Nagel)	(нога)	(nagà, nãgas)	(ongle)	
53.	name	+	+	-	+	*(e)nomn-
		(Name)	(имя)		(nom)	
54.	neck	+	-	-	-	
		(Nacken)				
55.	new	+	+	+	+	*new-
		(neu)	(новый)	(naũjas)	(neuf)	
56	night	+	+	+	+	*nok^{w}t-
		(Nacht)	(ночь)	(naktìs)	(nuit)	
57.	nose	+	+	+	+	*nas-
		(Nase)	(нос)	(nosis)	(nez)	
58.	not	+	+	+	+	*ne-
		(nicht)	(не)	(nè)	(non)	
59.	one	+	+	+	+	*oi̯-n-
		(ein)	(один)	(vìenas)	(un)	
60.	rain	+	-	+	-	*rek-
		(Regen)		(rõk-ti)		
61.	red	+	+	+	+	*reudh-
		(rot)	(рыжий)	(raũdas)	(rouge)	
62.	road	+	-	-	-	*reidh-
		(reiten)				
63.	root	+	-	-	+	*wərad-
		(Wurzel)			(racine)	
64.	sand	+	-	-	-	*sandh-(?)
		(Sand)				
65.	say, see	+	-	+	-	*sek^w-
		(sagen, sehen)		(saky´ti)		
66.	seed	+	+	+	+	*sē-(men)-
		(Same)	(семя)	(sḗmens)	(semence)	
67.	sit	+	+	+	+	*sed-
		(sitzen)	(сидеть)	(sēdḗti)	([as]-sis)	
68.	sleep	+	+	+	-	*sleb-
		(schlafen)	(слабый)	(slãbnas)		
69.	small	+	+	-	+	*(s)mal-
		(schmal)	(малый)		(mal)	
70.	smoke	+	+	+	-	*smeug-
		(schmauchen)	(смуглый)	(smáugti)		
71.	stand	+	+	+	+	*stā-
		(stehen)	(стоять)	(stó-ti)	(être)	
72.	star	+	-	-	+	*Haster-
		(Stern)			(étoile)	
73.	stone	+	+	-	-	*stei-
		(Stein)	(стена)			
74.	sun	+	+	+	+	*swel-
		(Sonne)	(солнце)	(sãulē)	(soleil)	

75.	swim	+ (schwimmen)	-	+ (sùmdyti)	-	*swem-
76.	tail	-	-	-	-	*dēḱ-
77.	that, this	+ (der)	+ (тот)	+ (tà-s)	+ (tel)	*to
78.	tongue	+ (Zunge)	+ (язык)	+ (liežùvis)	+ (langue)	*i̯enǵhu- / denǵhu-
79.	tooth	+ (Zahn)	-	+ (dantìs)	+ (dent)	*dent-
80.	tree	+ (treu, Teer)	+ (дерево)	+ (dervà)	-	*derw-
81.	two	+ (zwei)	+ (два)	+ (dù)	+ (deux)	*dwo(u)
82.	warm	+ (warm)	+ (гореть)	+ (gãras)	+ (four)	*g^{w}her-
83.	water	+ (Wasser)	+ (вода)	+ (vanduõ)	-	*wed-
84.	we	+ (wir)	-	+ (vè-du)	-	*we-
85.	who, what	+ (wer, was)	+ (кто, что)	+ (ka-s)	+ (que)	*k^{w}e-/*k^{w}i-
86.	white	+ (wei-)	+ (свет)	+ (šviẽsti)	+ (verre)	*ḱweit-
87.	woman	+ (Weib)	-	-	-	*weip-?
88.	yellow	+ (gelb)	+ (желтый)	+ (gel~tas)	+ (fiel)	*gh^{w}el-
89.	you	+ (ihr)	-	+ (jũs)	-	*i̯u-

Out of 89 roots from the English one-hundred word list[29] we have 87 correspondences with German (98%); 52 correspondences with Russian (58%); 51 correspondences with Lithuanian (57%) and 48 correspondences with French (54%).

Thus we obtain practically the same results as were obtained in our discussion of samples of English and Russian roots from arbitrarily chosen texts. This evidently shows that the distribution of individual frequency characteristics of roots in the Swadesh list coincides with their usual distribution in texts[30]. From this one can reach some important conclusions:

(a) the 'stability' of roots does not depend on the 'stability' of the words derived from them. This follows first of all from the third postulate of root glottochronology, but is also well demonstrated by the example of the Swadesh list discussed above: words included in it are intentionally more 'stable' than most

[29] Eight loans, *bark, big, dog, egg, mountain, person, round* and *skin* have been eliminated from the list. Pairs of words with similar roots are combined. These include *say, see; who, what; that, this.*

[30] The explanation of this is perhaps the fact that the Swadesh list includes the most common and the most usual ideas from quite different semantic fields, which thus really creates some kind of elementary conceptual text concerning humans and their environment. It is remarkable that many (22!) English roots from the Swadesh list are represented in the text discussed above: *full, what/ who, go, root, long, that/this, new, we, know, give, one, many, all, two, lie, not, small, tree, sit, hand, hear/show, I.*

words in a randomly chosen text, while the stability of roots is the same; (b) in the absence of texts (which unfortunately is often the case with many lesser-known languages) the Swadesh one-hundred word list could be used as a text for root glottochronology; (c) the mathematical apparatus worked out for classical glottochronology could be transferred to root glottochronology, but obviously requires another value for λ.

The third conclusion is the strongest, and will acquire additional practical experimentation. According to the preliminary results, however, the time calculated by inserting the proportion of root correspondences into the formula

$$(14)^{31}\ t = \sqrt{\frac{\ln (N(t)/ N_0)}{-\lambda N(t)}}$$

with λ = 0.035 corresponds in general with the datings obtained by the version of classical glottochronology proposed above with formula (13). Compare the dates of divergence of Russian from Polish, Lithuanian, German, and French (with time given in millennia):

	Polish[32]	Lithuanian[33]	German[34]	French[35]
t according to classical glottochronology	1.3	3.1	4.7	4.7
t according to root glottochronology	1.2	3.2	4.9	5.1

[31] Recently D. Leshshiner has suggested a different formula for dating language divergence within root glottochronology:

$$t = \frac{1 - \ln(N(t))}{1-\ln(1-N(t))} \times T$$ where T (the period of "halfdecay") is equal to 5.

[32] In Russian and Polish the following words from the 100 word list do not match: *život — brzuch, bol'šoj — wielki (duży), grud' — pierś, žeč' — palić, glaz — oko, žir — tšuszcz, xorošij — dobry, pečen' — wątroba, mnogo — wiele (dužo), luna — księżyc, rot — usta, krasnyj — czerwony, skazat' -powiedzieć (rzec), koža — skóra, xvost — ogon, ženščina — kobieta, xolodnyj — zimny.* Taking two borrowings in the Russian list into account *(oblako* and *sobaka)* we arrive at 85% matches.

[33] In Russian and Lithuanian the following words from the 100 word list match: *vesŭ- vìsas, kusat' — ką́sti, krov' — kraũjas, žeč'- dègti, nogoť— nãgas, xolodnyj — šaltas, pri-jti — ateĩti, umeret' — mirti, suxoj — saũsas, uxo — ausìs, zemlya —žẽme, ogon' — ugnìs, letet' — 1ẽkti, polnyj — pìlnas, dat' — dúoti, zelenyj - žãlias, ruka — rankà, golova — galvà, serdce — širdìs, rog — rãgas, ja — àš, znat' — žinóti, koleno — kẽlis, pečen' — kẽpenys, dlinnyj — ìlgas, mjaso — mēsà, novyj — naũjas, noč' — naktìs, nos — nósis, ne — nè, odin — víenas, sem'a — sḱla, sideť— sēdẽi, dym — dū́mai, stoyať— stovẽi, zvezda — žvaigždę̃ kamen' — akmuõ, solnce — sàulę̃plavat'— plaũkti, tot — tàs, ty — tù, jazyk — liežùvis, dva — dù,voda — vanduõ, my — mẽs, čto — kàs, belyj — báltas, želtyj — geltónas;* thus (not taking the words *oblako* and *sobaka* into account) exactly 50 % matches.

[34] In Russian and German the following words from the 100 word list form matches: *nogot' — Nagel, uxo — Ohr, est' — essen, jajco — Ei, pero — Feder, polnyj — voll, serdce — Herz, ja — ich, znat' — kennen, ležat' — liegen, dlinnyj — lang, mužčina — Mann, imja — Name, novyj — neu, noč' — Nacht, nos — Nase, ne — nicht, odin — ein, sem'a — Same, sidet' — sitzen, stoyat' — stehen, solnce — Sonne, etot — dieser, ty — du, jazyk — Zunge, dva — zwei, voda — Wasser, čto — was, kto — wer.* Taking into account two borrowings in the Russian list *(oblako, sobaka)* and three borrowings in the German list *(Fett, Kopf; rund)* we arrive at 30% matches.

[35] In Russian and French the following words from the 100 word list match: *kora — écorce, nogot'— ongle, umirat'— mourir, pit' — boire, uxo — oreille, jajco — oeuf, polnyj — plein, dat' — donner, serdce — coeur, ja — je, znat' — connaître, dlinnyj — long, luna — lune, imja — nom, novyj — neuf, noč' — nuit, nos — nez, ne — ne, odin — un, videt' — voir, sidet' — (être) assis, dym — fumée, stojat' — (être) debout, solnce — soleil, ty — tu, jazyk — langue, dva — deux, čto — que, kto — qui.* Taking into account two borrowings in the Russian list *(oblako, sobaka)* and three borrowings in the French list *(chemin, pierre, blanc)* we arrive again at 30 % matches.

Some of these datings are possibly too early. This is particularly true for the Russian-Polish divergence, where the increase in the percentage of correspondences could be due to secondary contacts: other Slavic languages allow us to date the Slavic divergence 200 to 300 years earlier. But in general they seem quite reasonable.

The method of 'root glottochronology' has some advantages over classical glottochronology. In particular, there are no problems with the choice of a 'main' word from several synonyms. It is also possible, to carry out a statistical analysis on a number of results, because the number of texts sampled — in contrast to the BL list — is in principle unlimited[36]. We can foresee, however, that etymostatistical analysis will face major obstacles when applied to languages whose history is less known. I would underline, however, that a thorough comparative historical analysis of the language data should precede any lexicostatistical or etymostatistical study, which will in any case be complementary to rather than a substitute for, comparative work.

Etymostatistics is currently being applied successfully to Semitic and Afro-Asiatic language materials by A. Militarev, and to Austronesian by Y. Sirk.

As we can see, the combination of lexicostatistics and etymostatistics allows us to obtain more precise datings and classifications, both for normal families and macro-families. According to preliminary results, modern Nostratic languages reveal levels of 15-20% correspondence according to root glottochronology, when one compares randomly-chosen texts with reliable dictionaries. These are much higher percentage values than those found between modern Nostratic languages on the 100-word Swadesh list (on average 5-9%), and I think that in the future, when we have a better understanding of comparative phonology and etymology, root glottochronology will play an important role in testing theories of remote relationships and in creating genetic classifications.

— Translated by N. Evans and I. Peiros.

TEXTS USED FOR ETYMOLOGICAL STATISTICS

(1) Aufrecht T. *Die Hymnen des Rigveda, I.*
Berlin 1995, pp. 1-2.

(2) Böll H. *Wanderer, Konunst du nach Spa... (Erzählungen).*
München 1971, pp. 7-8.

(3) Caesar G. Julius. *Commentarii de bello Gallico (liber primus).*
Moskva 1946, pp. 37-48.

(4) Čukovsky K. *Stixi i Skazki.*
Moskva 1984, pp. 37-39.

(5) Dauzat A. *Dictionnaire étymologique de la langue française.*
Paris 1938, pp. V-VI

(6) Freidenberg O.M. *Mif i literatura drevnosti.*
Moskva 1978, pp. 206-207.

(7) *Kulturnoe nasledie Vostoka (Problemy, poiski, suždenija).*
Leningrad 1985, pp. 34.

(8) Moloxovec E. — *Podarok molodym xoz'ajkam.*
Sankt-Peterburg 1904, p. 288.

[36] One further possible application of glottochronology is the dating of texts. This is however a separate area, which has its specific problems, and we will not discuss them here.

(9) *Sophoclis tragoediae.*
Moskva 1884, pp. 111-112.
(10) Russian original of this paper,
Moskva 1989.

LITERATURE

Arapov, M., Xerc, M. 1974: *Matematičeskie metody v istoričeskoj linguistike.*
Moskva.

Bergsland, K.,Vogt, H. 1962: *On the validity of glottlchronology.*
In Current Anthropology, v. 3.

Chretien, D. 1952: *The Mathematical Model of Glottochronology.*
In Language, v. 38.

Dyen, J., James, A. 1967: *English Divergence and Estimated Word Retention Rate.*
In Language, v. 47.

Embleton, S. 1986: *Statistics in Historical Linguistics.*
Bochum.

Fodor, J. 1961: *A glottochronologia ervenyessege a szlav nyelvek anyaga alapjan.*
In Nyelvtudomanyi Kozlemenyek, v. 63, No. 2.

Guryčeva, M. 1959: *Narodnaya latyn'*
Moskva.

Hattori Shiro: 1954: *"Gengo nendaiku" sunawachi "goito:keigaku" no ho:ho: ni tsuite (Nihon sogo no nendai).*
In Hattori Shiro:. Gengogaku no ho:ho:, Tokyo 1960, pp.515-566.

Hattori Shiro: 1959: *Nihongo no keito:*
Tokyo.

Helimski, E. 1984: *Problemy granic nostratičeskoj makrosem'ji jazykov.*
In Problemy izučenija nostratičeskoj makrosem'ji jazykov. Moskva, pp. 31-48.

Helimski, E. 1986: *K izučeniju i nadežnosti indoevropejsko-semitskikh leksičeskix sootvetstvij.*
In Balkany v kontekste Sredizemnomor'ja. Moskva.

Illič-Svityč, V. OSTJa 1971: *Opyt sravenija nostratičeskikh jazykov, t. 1,*
Moskva.

Merwe, N. van der, 1966: *New Mathematics for Glottochronology.*
In Current Anthropology, v. 7.

Muller, H. 1929**:** *Chronology of Vulgar Latin.*
In Zeitschrift für Romanische Philologie, Beiheft 78.

O'Neil, W. 1954: *Problems in the Lexicostatistic Time Depth of Modern Icelandic and Modern Faroese.*
In General Linguistics, v. VI, No. 1.

Swadesh, M. 1960: *Leksikostatističeskoe datirovanie doistoričeskikh etničeskikh kontaktov.*
In Novoe v lingvistike, vyp. 1. Moskva.

Swadesh, M. 1960a: *K voprosu o povyšenii točnosti v leksikostatističeskom datirovanii.*
In Novoe v lingvistike, vyp. 1. Moskva.

Swadesh, M. 1965: *Lingvističeskie sv'azi Ameriki i Evrazii.*
In Etimologja 1964, Moskva

Yakhontov, S. 1984: *Glottoxronologija: Trudnosti i perspektivy.*
In Drevnejšaja jazykovaja situacija v Vostočnoj Azii, pp. 39-47.