Glottochronology & lexicostatistics. Starostin's method…

Mikhail Vasilyev

Swadesh's glottochronology

Main principles (postulates)

1. In the lexicon of any language one can distinguish a set of words characterized by a particular stability. We will call this set "stable" or "basic".

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).

3. The proportion of words from BL which remains changeless (i.e. are not replaced by other words) over a certain time interval is constant. It depends only on the amount of time elapsed, and not on how that interval was chosen, or on which words and from what language were 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 any other daughter language.

(Arapov & Cherz 1974, Starostin 2000)

Basic equation of glottochronology

$$N(t) = N_0 \cdot e^{-\lambda t},$$

t – time period between two stages in the development of one and the same language (in millennia);

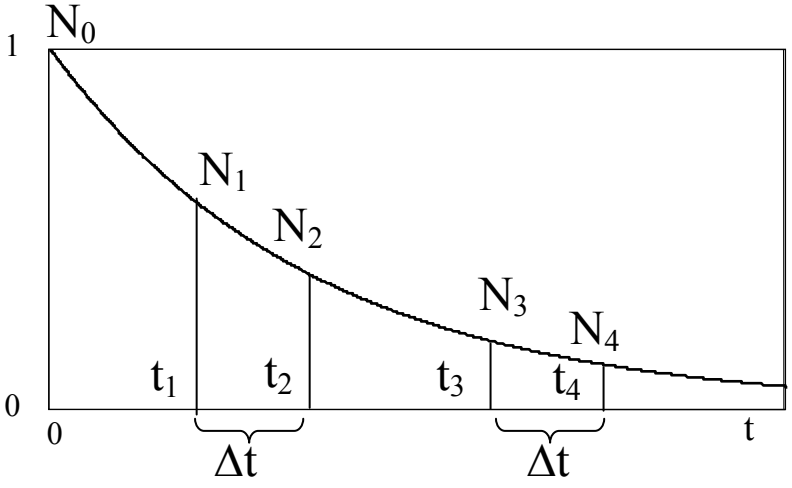
N_0 – set of words in the initial BL;

λ – rate of replacement (according to M.Swadesh λ=0,14);

$N(t)$ – proportion of wordlist items retained at the end of period t.

$$\frac{N_2}{N_1} = \frac{N_4}{N_3}$$

$$\lambda = -\frac{dN/N(t)}{dt} = \text{const}.$$



Graph 1. Correspondence between exponential correlation $N(t)=N_0 \cdot e^{-\lambda t}$ and the 3rd postulate

Starostin's approach

Adjusting the replacement rate

Values of the replacement rate λ obtained for different well-documented languages, applying formula $N(t)=N_0 \cdot e^{-\lambda t}$ to given quantities of *t* and *N(t)* (Starostin 2000)

Language	*t*	*N(t)*	λ
Japanese	1,2	0,93	0,06
Chinese	2,6	0,77	0,10
English	1,3	0,88	0,10
German	1,2	0,94	0,05
French	1,5	0,90	0,07
Spanish	1,5	0,91	0,06

Average value: $\lambda = 0{,}06$

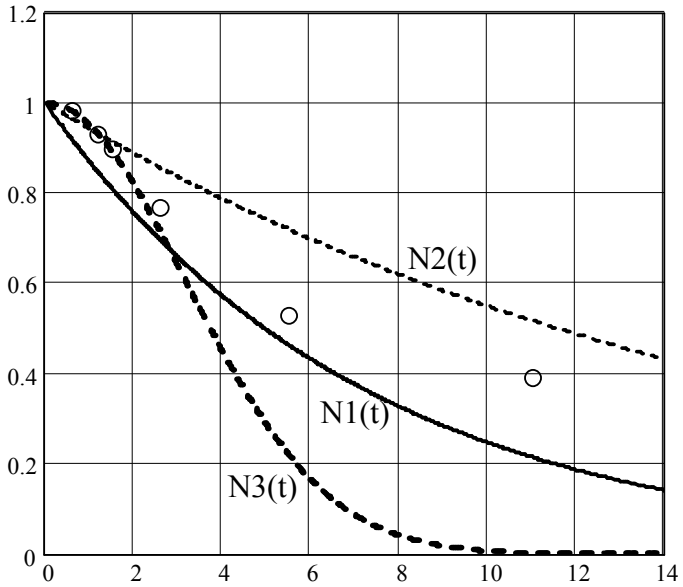
3rd postulate revision

Refutation: *the rate of change may differ from Swadesh's value of 0,14* (Bergsland and Vogt 1962, O'Neil 1964, Fodor 1961) *and depends on divergence time* (Starostin 2000).

Working hypothesis: *a word in contrast to a neutron can become 'older' and the probability of its retention in BL diminishes with the course of time. Thereby the rate of replacement is not a constant, but increases in direct proportion to time.*

$$\lambda(t) = \lambda \cdot t$$

Resulting formula: $N(t) = N_0 \cdot e^{-\lambda t^2}$



Graph 2. Comparative view of the models:

$N1(t) = e^{-0.14t}$ (Swadesh's formula, λ=0,14);

$N2(t) = e^{-0.06t}$ (λ-adjusting, λ=0,06);

$N3(t) = e^{-0.05t^2}$ (3rd postulate revision, $\lambda(t) = \lambda \cdot t$, λ=0,05)

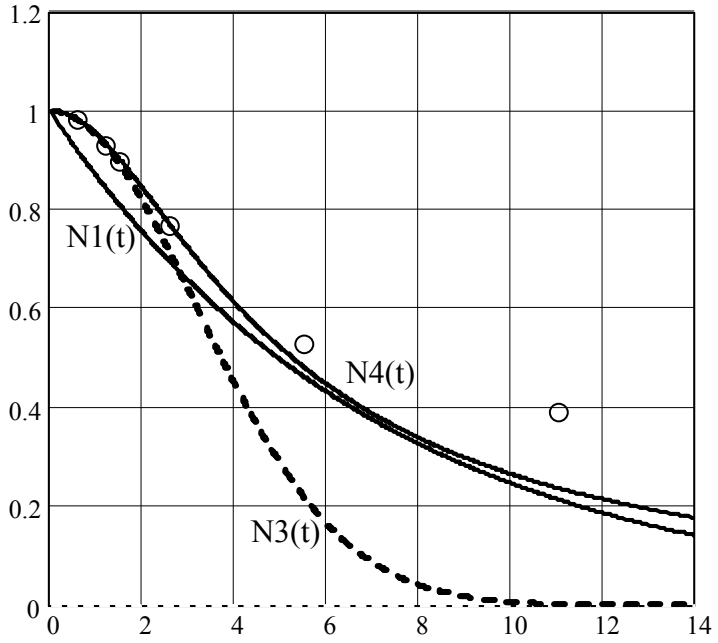
4th postulate revision

Refutation: *different items of the BL are not homogeneous by their stability and have different retention rates* (van der Merwe 1966, Dyen & James 1967).

Working hypothesis: *words should be replaced in turn, beginning with the least stable and going on to the more stable. As the most stable items progressively dominate in the wordlist, the average rate slows down in direct proportion to the percentage of retained words.*

$$\lambda(t) = \lambda \cdot N(t)$$

Resulting formula: $N(t) = N_0 \cdot e^{-\lambda N(t) t^2}$



Graph 3. Comparative view of the models:

$N1(t) = e^{-0.14t}$ (Swadesh's formula, λ=0,14);

$N3(t) = e^{-0.05t^2}$ (3rd postulate revision, $\lambda(t) = \lambda \cdot t$, λ=0,05)

$N4(t) = e^{-0.05 \cdot N4(t) \cdot t^2}$ (4th postulate revision, $\lambda(t) = \lambda \cdot t$, $\lambda(t) = \lambda \cdot N(t)$)

Some shortcomings of Starostin's approach

1. Poor accuracy of dating at great time depths (>4 millennia). (See curves N3(t) and N4(t) on Graph 3)
2. A 'contradictory' character of the lexical replacement process: relation $\lambda(t) = \lambda \cdot t$ reflects the acceleration of loss caused by the 'aging' of words, while relation $\lambda(t) = \lambda \cdot N(t)$ represents the opposite trend – slowing down of the replacement due to the gradual predominance of the most stable items in BL (Starostin 2000). This contradiction obviously cannot be explained by the nature of language development.

Alternative approach

Single language development

Basic assumption: *the process of lexical change can be described by a sum of several exponential components with different replacement rates, which correspond to several groups of items in BL with different stability.*

General view: $N(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t} + c_3 e^{-\lambda_3 t} + \ldots + c_i e^{-\lambda_i t} = \sum_i c_i e^{-\lambda_i t}$,

Initial model[*]*:* $N(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t} + c_3 e^{-\lambda_3 t}$

Method of calibration: least-squares approximation

$$\sum_{i=1}^{n} (c_1 e^{-\lambda_1 \cdot t_i} + c_2 e^{-\lambda_2 \cdot t_i} + c_3 e^{-\lambda_3 \cdot t_i} - N_i)^2 \rightarrow \min$$

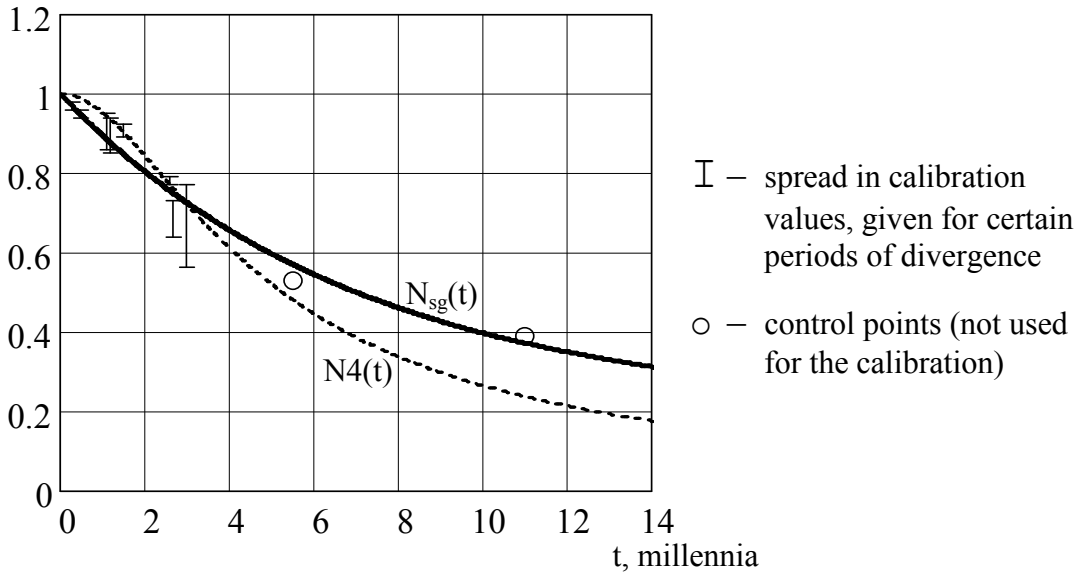
Calibration findings:

c_1=0,195 λ_1=0,000 – the most stable part of BL ('core' vocabulary)

c_2=0,199 λ_2=0,140 } since $\lambda_2 \approx \lambda_3$, these two components can be replaced by the

c_3=0,606 λ_3=0,146 } one with $c=c_2+c_3=0,8$ and $\lambda=0,14$

Resulting formula: $N_{sg}(t) = 0,2 + 0,8\, e^{-0,14 \cdot t}$



Graph 4. Models of a single language development

$N4(t) = e^{-0.05 \cdot N4(t) \cdot t^2}$ (Starostin's approach)

$N_{sg}(t) = 0,2 + 0,8\, e^{-0,14 \cdot t}$ (the proposed approach)

[*] According to Khinchin theorem a sum of independent streams of random events comparable by their rate can be substituted by a certain single stream (Khinchin 1963). Thus, even if each meaning in BL has its own rate (following e.g. Dyen & James 1967), we can expect the number of components to be reduced to only a few during the calibration.

Relative divergence of two daughter languages

According to the 5th postulate

Swadesh's formula: $N_{2Sw}(t)=N_A(t)\cdot N_B(t)=N(t)\cdot N(t)=e^{-\lambda t}\cdot e^{-\lambda t}=e^{-2\lambda t}$

Starostin's formula: $N_{2St}(t)=e^{-2\cdot 0,05\sqrt{N_2(t)}\cdot t^2}$

5th postulate revision

Refutation: *datings of divergence obtained by comparison of two daughter languages BLs are much younger then those obtained for the percent of cognates between one of these languages and their proto-language* (see e.g. Fodor 1961). *This can be explained only if we allow a consistency in separate development of daughter languages after the disintegration of the proto-language* (Arapov & Cherz 1974, an illustration of this consistency provided in Appendix 1 below).

Basic assumption: *in the word list of any daughter language one may distinguish a group of items, that develops 'coherently' in all daughter languages and diminishes in the course of time. Therefore, the process of the relative divergence can be represented by two components one of which reflects the 'coherent' part of BL and the other – its independent part.*

General view: $N_{rel}(t)=N_c(t)+N_i(t)$

Coherent component: $-dN_c=\mu\cdot N_c(t)\cdot dt$

Independent component: $N_i(t)=N_{rel}(t)-N_c(t)$, $-dN_i=\eta\cdot N_i(t)\cdot dt$

Resulting equation: $\frac{1}{\eta}\cdot\frac{dN_i}{dt}+N(t)=N_0\cdot e^{-\mu t}$

General solution: $N_{rel}(t)=N_0\left(\frac{\mu}{\mu-\eta}\cdot e^{-\eta t}+\frac{\eta}{\eta-\mu}\cdot e^{-\mu t}\right)$

*Initial model***:* $N_{rel}(t)=c_0+c_1\left(\frac{\mu}{\mu-\eta}\cdot e^{-\eta t}+\frac{\eta}{\eta-\mu}\cdot e^{-\mu t}\right)$

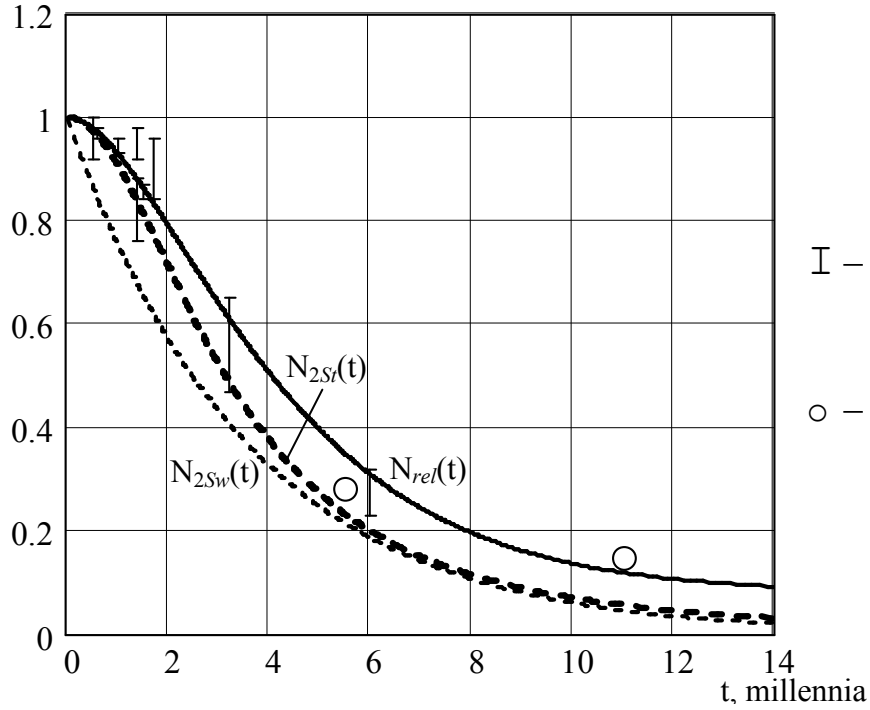
*Final simplified formula****:*

$$N_{rel}(t)=0,08+0,92\cdot e^{-0,45\,t}(1+0,45\,t)$$

(for the function table see Appendix 2 below)

** The constant c_0 was introduced to allow for a possible coincidence between the 'core' vocabularies of the daughter languages (in consideration of the results obtained for the model of a single language development $N_{sg}(t)$ (above)).

*** During the calibration μ and η turned out to be approximately equal ($0,457\approx 0,454$), which enabled to simplify the initial expression to this form: $N_{rel}(t)=c_0+c_1\cdot e^{-\eta t}(1+\eta\cdot t)$, $(\mu=\eta)$.

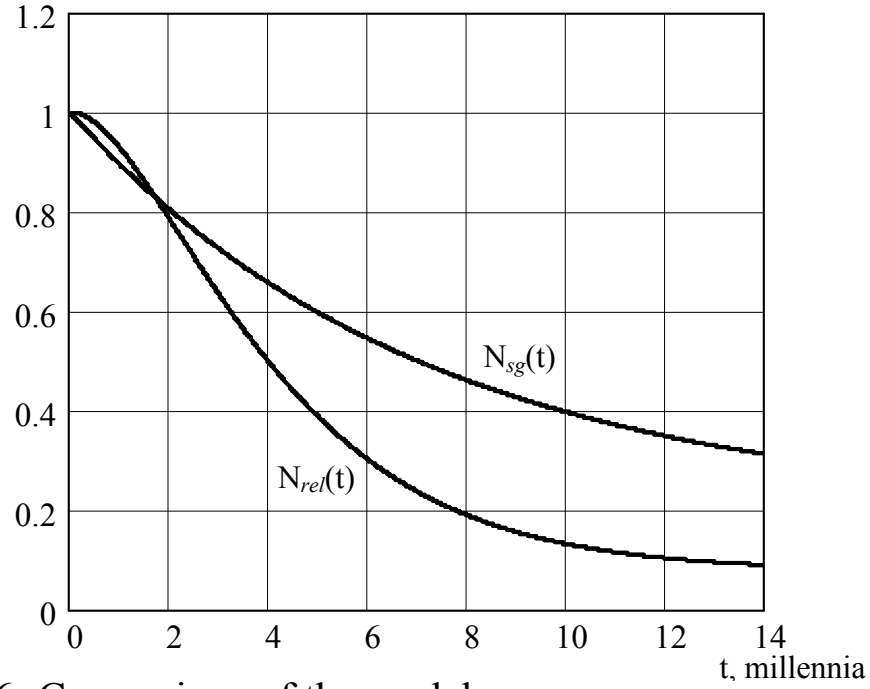


Graph 5. Models of two daughter languages relative development

$$N_{2Sw}(t) = e^{-2 \cdot 0.14\,t} \quad \text{(Swadesh's formula)}$$

$$N_{2St}(t) = e^{-2 \cdot 0{,}05 \sqrt{N_2(t)} \cdot t^2} \quad \text{(Starostin's formula)}$$

$$N_{rel}(t) = 0{,}08 + 0{,}92 \cdot e^{-0{,}45t}(1 + 0{,}45t) \quad \text{(the new approach)}$$



Graph 6. Comparison of the models:

$$N_{sg}(t) = 0{,}2 + 0{,}8\,e^{-0{,}14 \cdot t} \quad \text{(single language development)}$$

$$N_{rel}(t) = 0{,}08 + 0{,}92 \cdot e^{-0{,}45t}(1 + 0{,}45t)$$

(two daughter languages development)

Some conclusions

1. The proposed attempts to improve Swadesh's formula by applying different additional correlations to the rate of change (λ) were shown to be inefficient in obtaining accurate datings for the whole period of time.
2. The variation of the divergence speed should be explained not by a change of the average replacement rate with the course of time, but due to the existence in BL several groups of items with different, but constant stability rates.
3. During the calibration one of these groups was found to have a particularly low rate of change close or equal to zero. The meanings from this group (≈ 20 items) constitute the most stable and nearly invariable part of BL, so-called 'core' vocabulary. This constant component together with an exponential component, whose rate is equal to 0,14 , are represented by the revised glottochronological formula:

$$N_{sg}(t) = 0{,}2 + 0{,}8\,e^{-0{,}14\cdot t}$$

4. The refutation of the fifth postulate makes it incorrect to use the model of a single language development for dating the divergence between two related languages.
5. The revealed consistency in the development of daughter languages was taken into account and given a formal expression (as the 'coherent' component $N_c(t)$) in the proposed model of the relative divergence between two languages:

$$N_{rel}(t) = 0{,}08 + 0{,}92 \cdot e^{-0{,}45t}(1 + 0{,}45t)\,.$$

Appendix 1

Identical replacements occurred in the BLs of seven modern Korean dialects during their separate development from the common ancestor (Middle Korean) over a period of 500 years

Meaning	Middle Korean AKO (1500 AD)	Chensando CHN	Phyenyang-namdo PNM	Kyensando KJN	Hamgyengdo HMG	Chejudo CJD	Seoul SEU	Kanwendo KNW
'feather'	čís	thɔl	thɔl	"	thɔl	"	thɔl	thɨl
'hair'	thə̀rí, thə̀rə̀k	khal	k̄al	khál	khâl	–	qhal	"
'knee'	mù̀rù̀p(h), mù̀rà̀p	"	"	ćêŋgej	"	"	"	oɣumpẽ
'full'	kằtằk-/kằtằik-	"	"	"	ćhâù̀da	"	čäwúda	"
'head'	mə̀rí	"	"	"	"	kol	"	kol
'skin'	kà̀čò̀k, kà̀čh	"	"	"	"	koptegi	"	"

" – the word remains unchanged or has been replaced by a borrowing.

Appendix 2

Values of the divergence time *t* obtained by the model $N_{rel}(t) = 0{,}08 + 0{,}92\,e^{-0{,}45t}(1 + 0{,}45\,t)$ for the range of N(t) from 99 to 10 words retained in BL

N_{rel}(t)	t
1	0,00
0,99	0,34
0,98	0,50
0,97	0,62
0,96	0,73
0,95	0,83
0,94	0,92
0,93	1,01
0,92	1,09
0,91	1,17
0,9	1,24
0,89	1,32
0,88	1,39
0,87	1,46
0,86	1,53
0,85	1,60
0,84	1,67
0,83	1,74
0,82	1,81
0,81	1,87
0,8	1,94
0,79	2,00
0,78	2,07
0,77	2,14
0,76	2,20
0,75	2,27
0,74	2,33
0,73	2,40
0,72	2,46
0,71	2,53

N_{rel}(t)	t
0,7	2,60
0,69	2,66
0,68	2,73
0,67	2,80
0,66	2,87
0,65	2,93
0,64	3,00
0,63	3,07
0,62	3,14
0,61	3,21
0,6	3,28
0,59	3,36
0,58	3,43
0,57	3,50
0,56	3,58
0,55	3,65
0,54	3,73
0,53	3,81
0,52	3,89
0,51	3,97
0,5	4,05
0,49	4,13
0,48	4,21
0,47	4,30
0,46	4,39
0,45	4,48
0,44	4,57
0,43	4,66
0,42	4,75
0,41	4,85

N_{rel}(t)	t
0,4	4,95
0,39	5,05
0,38	5,16
0,37	5,26
0,36	5,38
0,35	5,49
0,34	5,61
0,33	5,73
0,32	5,85
0,31	5,98
0,3	6,12
0,29	6,26
0,28	6,41
0,27	6,56
0,26	6,72
0,25	6,89
0,24	7,07
0,23	7,25
0,22	7,45
0,21	7,67
0,2	7,89
0,19	8,14
0,18	8,41
0,17	8,71
0,16	9,03
0,15	9,40
0,14	9,82
0,13	10,32
0,12	10,92
0,11	11,68
0,1	12,75

References

Arapov & Cherz 1974 — Arapov, M. & M. Cherz. *Matematičeskie metody v istoričeskoj linguistike.* Moscow: Nauka.

Bergsland & Vogt 1962 — Bergsland, K. & H.Vogt. *On the validity of glottochronology.* Current Anthropology 3, pp.111-153.

Dyen & James 1967 — Dyen, J. & A. James. *English divergence and estimated word retention rate.* Language 47.

Fodor 1961 — Fodor, J. *A glottochronologia ervenyessege a szlav nyelvek anyaga alapjan.* Nyelvtudomanyi Kozlemenyek 63(2).

van der Merwe 1966 — Merwe, N. van der. *New mathematics for glottochronology.* Current Anthropology 7, pp. 485-500.

O'Neil 1964 — O'Neil, W. Problems in the lexicostatistic time depth *of modern Icelandic and modern Faroese.* General Linguistics VI(1).

Swadesh 1960a — Swadesh, M. *Leksikostatističeskoe datirovanie doistoričeskikh etničeskikh kontaktov.* Novoe v lingvistike 1.

Swadesh 1960b — Swadesh, M. *K voprosu o povyšenii točnosti v leksikostatističeskom datirovanii.* Novoe v lingvistike 1.

Starostin 2000 — Starostin, S. *Comparative-Historical Linguistics and Lexicostatistics.* // Time Depth in Historical Linguistics, ed. by Colin Renfrew, April McMahon & Larry Trask. McDonald Institute for Archaeological Research, Cambridge, pp. 223–259.

Khinchin 1963 — Khinchin, A. *Raboty po matematičeskoj teorii massovogo obsluživanija.* Moscow: Fizmatlit.