BALTISTICA XLII(2) 2007 185–210

Petra NOVOTNÁ, Václav BLAŽEK
Masaryk University

GLOTTOCHRONOLOGY AND ITS APPLICATION TO THE BALTO-SLAVIC LANGUAGES

In memoriam of Sergei Starostin
(March 24, 1953 – Sept. 30, 2005)

The explicit purpose of this contribution is to present a quantitative approach to the genetic classification of the Balto-Slavic languages. The implicit aim represents an attempt to rehabilitate the method called 'glottochronology'. Although the method developed by Morris Swadesh was rightfully criticized by specialists in the Indo-European languages, this does not mean that it is impossible to reconstruct the processes of divergence of related languages including their absolute chronology. The radical modification of the 'classical glottochronology' formulated by Sergei Starostin (1989; 1999) eliminates its most egregious mistakes and gives a tool for quite realistic estimates of an absolute date. The present article should serve as an illustration, which is in good agreement with both the data of archaeology and historical facts as well. The last, but not least reason for this topic is to mention the scientific heritage of Sergei Starostin, an excellent linguist and great man, who left us so unexpectably, but did so much.

0. Radiocarbon method.
1. 'Classical glottochronology' according to Swadesh.
2. 'Recalibrated glottochronology' according to Starostin.
3. Lexicostatistics and glottochronology applied to Slavic languages.
4. Lexicostatistics and glottochronology applied to Baltic and Balto-Slavic languages.
5. Correlations with the extralinguistc disciplines: history and archaeology.
6. Conclusion.

0. The method called *glottochronology* represents an attempt to date the divergence of related languages in absolute chronology. Its author, Morris Swadesh, was inspired by another method, used for dating organic remnants, the so called radiocarbon method. Let us repeat the main steps in the deduction of the method. In the beginning it was the discovery of the radiocarbon isotope C^{14}, existing in the atmosphere in the proportion 1 : 10^{12} with the usual isotope C^{12}. Thanks to the food-chain, the radioactive isotope occurs in green plants and consequently in biological tissues of animals. After the death of any living organism the disintegration of the radioactive isotopes according to the exponential function follows. The exponential disintegration means that after the constant time period T (= half-time of disintegration) the concentration of the radioactive isotope falls in a half, after 2T in a quarter, etc. On the basis of this phenomenon, W. F. Libby developed the radiocarbon method (1947), serving to determine the age of organic remnants younger than 50 millennia. The method was recently defined with more precision (e.g. the change of the half-time from 5568 to 5730 years; correlation with dendrochronology, etc.), but its basic idea remains. Regarding the fact that M. Swadesh borrowed the mathematic apparatus from Libby, it is useful to repeat it.

(1) $\Delta N(t) = -\lambda \cdot N(t) \cdot \Delta t$ ... decrease ΔN from N radioactive nuclei in the time interval Δt, where λ is a constant of proportion

(2) $dN(t) = -\lambda \cdot N(t) \cdot dt$ ... approximation of discrete quantities by connected ones, allowing the integration

$$\int \frac{dN(t)}{N(t)} = \int -\lambda \cdot dt$$ ... leading to the solution

$\ln N(t) = -\lambda \cdot t + C$. After delogarithmization we reach

$N(t) = e^{-\lambda t + C} = e^{-\lambda t} \cdot e^{C}$, where $e^{C} = K$. So we can write

$N(t) = K \cdot e^{-\lambda t}$.

It remains to determine the function of the constant K. It is possible thanks to the initial conditions, i.e. in the time $t = 0$, when $N(t) = N_0$:

(3) $N(t) = N_0 \cdot e^{-\lambda t}$, where N_0 represents the number of undisintegrated nuclei at the beginning of the process.

From the equation (3), which is a standard solution of the differential equation (2), we deduce the significance of the *half-time of disintegration* T,

defined as the time interval, in which the number of the undisintegrated nuclei decrease in $^1/_2$:

(4) $N(T) = {}^1/_2\, N_0$

$^1/_2\, N_0 = N_0 \cdot e^{-\lambda T}$, after a reduction

$^1/_2 = e^{-\lambda T}$, after logarithmization

$\ln {}^1/_2 = -\lambda T$, i.e. $\ln 2 = \lambda T$, or

(5) $T = \frac{\ln 2}{\lambda}$

The half-time of disintegration of the radioactive isotope C^{14} was empirically established as 5730 years. It allows one to determine the value of the constant of disintegration λ.

For practical calculations it is helpful to use the formula, derived from the definition of the half-time of disintegration. If the number of the undisintegrated nuclei decreases in $^1/_2$ after every time period T, we get:

(6) $N(t) = N_0 \cdot (\frac{1}{2})^n$, where n means, how many periods T correspond with the age of the specimen. Hence

$\frac{N(t)}{N_0} = (\frac{1}{2})^n$, i.e. $\frac{N_0}{N(t)} = 2^n$. Let us logarithmize it:

$\ln \frac{N_0}{N(t)} = \ln 2^n = n \cdot \ln 2$ and we reach

(7) $n = n = \frac{\ln \frac{N_0}{N(t)}}{\ln 2}$

From here we get the age of the specimen

(8) $t = n \cdot T$.

1. Around 1950 Libby's radiocarbon method inspired one American anthropologist and specialist in native American languages, Morris Swadesh, to extend its application to the development of languages. His goal was the absolute dating of the time of divergence of related languages. Swadesh thought that the replacement of words in languages is determined by exponential rule similar to the disintegration of radioactive nuclei of isotope C^{14}. He needed to calculate the rate of this change. For this reason he established a testing word-list, consisting first of 215, later of 200 semantic units, which had to be universal and immune from borrowing. Thanks to the cooperation of specialists in sinology, egyptology, classical philology, Romance and Germanic linguistics, he was able to determine

the average constant of disintegration applied to one millennium, in 19,5% changes in the testing word-list, i.e. on average 80,5% of the units of the basic word lexicon in the development of one language should be preserved during this period (see Swadesh 1952). Naturally, if the constant is really universal. In 1955 Swadesh published a new study, reflecting the first critical reactions. He radically reduced and changed the testing word-list. The new list consisted of 100 semantic units. On the basis of the reduced 'basic lexicon', the constant of disintegration was changed to 14% per. millennium, i.e. 86% of the lexical units should be preserved in the development of one language after one millennium. The elementary postulates may be formulated as follows:

[1] In the lexicon of every natural language it is possible to determine the part, which is more stable than others. Let us call it the *basic lexicon*.

[2] It is possible to define the set of meanings, expressed in every language by words from the *basic lexicon*. Let us designate it the *basic testing list* (BTL). The symbol N_0 will signify the number of various meanings, contained in the list.

[3] The share r of the words from the basic testing list preserved after the constant period Δt, is constant; i.e. it depends only on the length of the time interval, not on a concrete language or a choice of words.

[4] All words representing the basic testing list have equal chances of being preserved during the same time interval.

[5] The probability of being preserved for any unit from the basic testing list does not depend on the probability of being preserved in the basic testing list of another language.

To calculate the time passed between the existence of two languages A and B, where B is a descendant of A, Swadesh used the mathematical apparatus from the radiocarbon method. He began from equation (3):

(9) $N(t) = N_0 \cdot e^{-lt}$, where λ represents the analogy to the constant of disintegration in the equation (3). Exactly it is defined as the share of the words in the basic testing list, which are replaced during one millennium. Hence:

(10) $\frac{N(t)}{N_0} = e^{-\lambda t}$, or $\ln \frac{N(t)}{N_0} = -\lambda t$. From here

(11) $t = \frac{\ln \frac{N(t)}{N_0}}{-\lambda}$, or $\ln \frac{\ln c}{-\lambda}$, where $c = \frac{N(t)}{N_0}$.

If the share r from the postulate (3) is also related to the period of one millennium, it will represent the constant which is complementary to λ , i.e.

(12) $r = 1 - \lambda$.

For the decrease of the words from BTS per millennium the equation $\Delta N = N_0 - N(t_1) = N_0 - N_0 \cdot e^{-\lambda \cdot 1} = N_0(1 - e^{-\lambda})$ is valid. The same value must be reflected in the product $N_0 \cdot \lambda$. From the comparison $1 - e^{-\lambda 1} = \lambda = 1 - r$ (see 11) we reach

(13) $r = e^{-\lambda}$.

The same result is accessible from the comparison of the right sides of the equations expressing the shares of the preserved words in the BTL per millennium: $N = N_0 \cdot e^{-\lambda \cdot 1}$ & $N = N_0 \cdot r$.

Consequently it is possible to rewrite the equation (10) by means of (13) in the form

(14) $c = r^t$, where t indicates the time in millennia.

Regarding the postulate (5) the share c_2 of the preserved lexicon from the BTL in two related languages, i.e. the languages, developed from a common protolanguage, equal to the square of the share of the words preserved in the individual development:

(15) $c_2 ,= (r^t)^2 = r^{2t}$. Logarithmizing it, we express t:
$\ln c_2 ,= \ln r^{2t} = 2t \ln r$. From here

(16) $t = \frac{\ln c_2}{2 \ln r}$ or with respect to the equation (13)

(17) $t = \frac{\ln c_2}{-2\lambda}$,

where c_2 means the share of commonly inherited pairs of the words in BTL in both analyzed languages.

In application of glottochronology the formulae (16) or (17) are used most frequently. For illustration of the practical procedure let us to estimate the time of divergence of German and French. In the BTL of both languages there are 33 pairs of commonly inherited words. Both lists are complete, which means that $c_2 = 0{,}33$. Applying it for the equations (16) or (17), we reach the time of divergence in millennia:

(16) $t = \frac{\ln 0{,}33}{2 \ln 0{,}86} = \frac{-1{,}10866}{-0{,}30164} = 3{,}675$

It is more advantegous to calculate a rich set of data with corresponding share of preservation of BTL for one language (c_1) or for two related languages (c_2) – see table 1:

Table 1

c_1	0,99	0,97	0,95	0,90	0,85	0,80	0,75	0,70	0,65	0,60	0,55	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10
c_2	0,97	0,94	0,90	0,81	0,72	0,64	0,56	0,49	0,42	0,36	0,30	0,25	0,20	0,16	0,12	0,09	0,06	0,04	0,02	0,01
t	0,03	0,20	0,35	0,70	1,10	1,50	1,90	2,40	2,90	3,40	4,00	4,60	5,30	6,10	7,00	8,00	9,30	10,7	13,0	15,3

The time of divergence for German and French occurs in the line for t, corresponding with $c_2 = 0{,}33$. This value may be approximated between the times 3,40 a 4,00 millennia in table 1. Concretely it is possible to estimate the age of the common ancestor for German and French as 3700 BP or 1700 BC according to the methodology developed by Swadesh.

The preceding steps operated only with a pair of synchronic languages. It is also necessary to solve the situation, if each of the compared languages was recorded at a different time. Let us designate t_1 and t_2 the times from the disintegration of the common ancestor of the compared languages to their record in various times. In this case the equation (16) can be modified as $2t = \frac{\ln c_2}{\ln r}$, and further

(18) $t_1 + t_2 = \frac{\ln c_2}{\ln r}$.

Since t_1 and t_2 are usually unknown, only their subtraction Δt_{12} is at our disposal, it is possible to substitute the sum $t_1 + t_2$ by $t_1 + t_1 + \Delta t_{12} = 2t_1 + \Delta t_{12}$, where t_1 is shorter from both intervals t_1, t_2. From here for two asynchronically attested languages the final formula appears as follows:

(19) $t_1 = \frac{\ln c}{2 \ln r} - \frac{\Delta t_{12}}{2}$, where $t_1 = \min (t_1, t_2)$.

2. Swadesh's glottochronology was welcomed by specialists studying languages without a longer literary history. On the other hand, the sharpest negative reaction was from specialists in the Indo-European languages. This was understandable: the comparison of the glottochronological estimates with safely known facts from the known history of some Indo-European languages frequently indicated a big disagreement. More interesting than the aprioristic rejection was the criticism of the concrete premises, postulates, conclusions, especially, if the critics offered their alternative solutions. The most remarkable modifications eliminating some of the weak points of the method were formulated by the Canadian Sheila Embleton (1986) and the Russian Sergei Starostin (1989, English 1999). Both scholars agreed that the 'classical glottochronology' of Swadesh was mistaken in that the replacement of words was not distinguished from borrowing. E.g. such innovation was Russian *glaz*

"eye", which replaced common Slavic **oko*. On the other hand, it is possible to identify a borrowing, probably of Iranian origin, in Russian *sobaka* "dog", besides the less frequent *pës*, which reflects common Slavic **pъsъ* "dog". Starostin offered a simple solution: eliminate all borrowings before any calculation. Applying this procedure to the testing languages, used for the estimation of the constant of disintegration λ, we reach lower value of the constant and its significantly smaller dispersion (table 3).

Starostin compared the proportions of the inherited lexicon in histories of the same languages during various time of divergence, related to one millennium times, concretely in some Romance languages versus Vulgar Latin from the middle of the first mill. AD and versus early classical Latin from the time of Plautus, c. 200 BC. The values of c in the table 2 are calculated now without loans; time is expressed in millennia:

Table 2

TABLE 2 language	$c = \underline{N(t)}$, t = 1,5 No	$\lambda = \underline{\ln c}$, t = 1,5 t	$c = \underline{N(t)}$, t = 2,2 -t	$\lambda = \underline{\ln c}$, t = 2,2 -t
French	88/99 = 0,89	0,07	75/97 = 0,77	0,12
Spanish	90/98 = 0,92	0,06	79/97 = 0,80	0,10
Rumunian	87/96 = 0,91	0,06	76/95 = 0,80	0,10

For the differences between the results in the third and fifth columns Starostin finds the only explanation, the formula (11), implying $\lambda = \frac{\ln c}{-t}$ is not valid.

The empirical figures from the table 2 confirm that the optimal approximation is the function

$\lambda^* = \frac{\lambda}{t} = \frac{\ln c}{-t^2}$ (20).

The preceding thoughts are based on the data in the table 3.

Table 3

language	age t [millennia]	λ after Swadesh	λ without loans	$\lambda^* = \lambda / t$
English	1,3	0,14	0,10	0,08
German	1,2	0,08	0,05	0,04
Norwegish (riksmal)	1,0	0,20	0,05	0,05
Icelandic	1,0	0,06	0,06	0,06
French	1,5	0,09	0,07	0,05
Spanish	1,5	0,07	0,06	0,04
Rumunian	1,5	0,09	0,06	0,04
Japanese	1,2	0,11	0,06	0,05
Chinese	2,6	0,10	0,10	0,04

It is apparent that the dispersion of the 'constant of disintegration' λ according to Swadesh is very high, from 6 do 20%. After the elimination of borrowings, the dispersion of this value for the analyzed nine languages tapers to 5–10%. Still narrower will be the interval in the case, if λ is a function of time. Abstracting from rather specific English, the value oscillates from 4 to 6%. These results led Starostin to the new value of the 'constant of decrease': λ = 0.05 per millennium. The situation of English is more complex. It seems its development is faster than is usual in other languages. This phenomenon is undoubtedly connected with the massive influence of Old Norse in the period 800–1100 and Old French in the following five centuries, causing according to Starostin certain pidgin-like features in English. But even the new value of λ = 5% does not defend against tendency to reach a more recent date of divergence, especially in the case of longer time periods. Starostin seeks a solution in the following idea. It is empirically proven that individual words in the lexicon of every language, including BTL, are replaced unevenly. If the words in any language were ordered from least stable to most stable, the words with the lowest stability would be replaced most quickly, while the more stable words would have a longer life. This means, the speed of changes decreases over time. Summing up, "c" is not a constant, but a function of time, c = c(t) and formula (9) should be modified as follows:

(21) $N(t) = N_0 \cdot e^{-\lambda \cdot c(t) \cdot t^2}$ for a development of one language, where $c(t) = \frac{N(t)}{N_0}$, and

(22) $N(t) = N_0 \cdot e^{-2\lambda} \cdot \sqrt{c(t)} \cdot t^2$ for the divergence of two languages, developed from a common protolanguage.

From here it is possible to deduce for the time of development of one language (23), or for the time of divergence of two languages (24):

(23) $t = \frac{\sqrt{\ln c}}{\sqrt{-\lambda c}}$

(24) $t = \frac{\sqrt{\ln c}}{\sqrt{-2\lambda\sqrt{c}}}$

The result is a transcendental function, since c = c(t). The easiest way of determining of the time of divergence for the empirically investigated values is offered in table 4, calculated by Sergei Starostin:

Table 4

c_1	0,99	0,97	0,95	0,90	0,85	0,80	0,75	0,70	0,65	0,60	0,55	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10
c_2	0,97	0,94	0,90	0,81	0,72	0,64	0,56	0,49	0,42	0,36	0,30	0,25	0,20	0,16	0,12	0,09	0,06	0,04	0,02	0,01
t	0,3	0,8	1,0	1,5	2,0	2,4	2,8	3,2	3,7	4,1	4,7	5,3	6,0	6,8	7,8	9,0	10,7	12,7	16,6	21,5

Now it is possible to return to the question of the time of divergence between German and French. In both languages there are 3 loans in the BTL and 33 common cognates.

Hence

$$c_2 = \frac{33}{100-3-3} = \frac{33}{94} = 0{,}351 = 35{,}1\% \ .$$

The corresponding time of divergence is c. 4 220 years. Naturally, it is an exaggeration to conclude that two languages were separated in a single concrete decade. Better is to use the formulation that their common protolanguage disintegrated in the 23rd cent. BC.

2.1. The situation of two asynchronically attested languages is solved by Starostin differently from Swadesh. Starostin's strategy consists in projection of the historical data to the present level and only after this synchronization the same approach as for living languages is applied to them. It is useful to demonstrate this procedure on concrete idioms, e.g. classical Latin e.g. of Caesar (1st cent. BC) and Gothic of Wulfila's translation of the New Testament (4th cent. AD). The Latin corpus (i.e. the 100-word-list) is complete, while in the Gothic list 18 units are missing (if Crimean Gothic *ada* "egg" is included). This means, there are 82 common semantic pairs from the BTL and from them 39 cognates, i.e. etymologically related forms inherited from a common protolanguage. The proportion 39/82 means 47,6%. A language recorded at the time interval Δt ago would preserve till the present c-times less words from BTL. For Latin recorded 20.5 cent. ago it is c. 0.845. If Gothic would exist till the present time, in its hypothetical descendant the share of the preserved BTL would be 0.892 (see table 4). The common protolanguage of Latin and Gothic projected into the present would preserve $c_{LG} \cdot c_L \cdot c_G = 0.476 \cdot 0.842 \cdot 0.892 = 0.357$, i.e. 35,7% common words. Let us mention, the result of the comparison of German and French gave the share 0.351. This means, the dating of the divergence of the representatives of modern Germanic and

Romance languages is practically the same as the dating of the divergence of Latin and Gothic, the 23rd cent. BC. It seems to be natural, but for the 'classical glottochronology' it was an unattainable goal.

3. For the Slavic languages, quantitative methods as lexicostatistics or glottochronology were applied by various scholars. Let us begin with the attempts based on standard Swadesh's variant.

3.1.1. One of the most detailed attempts to apply 'classical glottochronology' for the Slavic languages is from Czech slavicists A. Lamprecht & M. Čejka (1963) and Čejka himself (1972). In his study from 1972 Čejka compiled the 100-word-lists from 12 living languages. His results are concentrated in the table 5 (the figures are %):

Table 5

	Mac.	SC.	Sln.	Slk.	Cz.	ULus.	LLus.	Pol.	Blr.	Ukr.	Rus.
Bul.	86	80	76	75	74	73	71	74	77	72	74
Mac.		84	75	76	75	76	73	71	74	71	70
SC.			85	80	79	77	74	75	77	73	71
Sln.				80	84	78	78	79	76	71	74
Slk.					92	86	87	85	80	76	74
Cz.						87	87	81	77	73	74
ULus.							94	80	78	74	74
LLus.								83	78	74	73
Pol.									80	76	77
Blr.										92	86
Ukr.											86

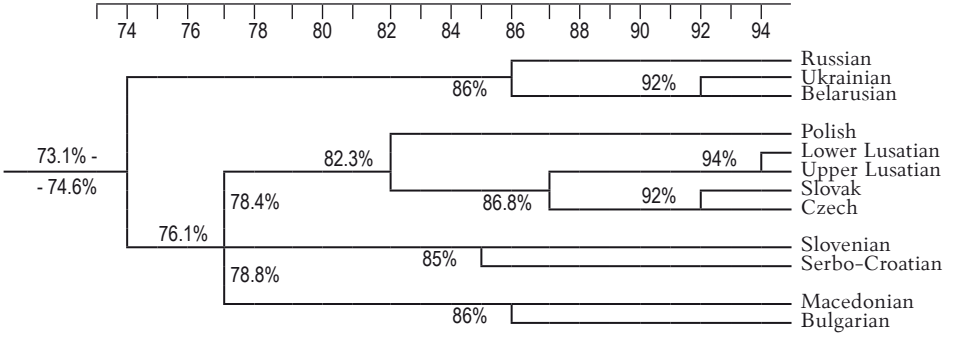
The following step consists in the determination of the closest pairs or groups of languages. The pairs (or triads etc.) with the highest grade of relationship will serve as the base of comparison, leading to the deeper past. The order of the first closest pairs is: ULus. + LLus. (= Lus.) 94%, Cz.+ Slk. (= Czsl.) 92%, Blr.+ Ukr. 92%, Rus. + [Blr. + Ukr.] (= ESl.) 86%, Bul. + Mac. 86%, SC. + Sln. 85%.

Table 6

	SC. + Sln.	Czsl.	Lus.	Pol.	ESl.
Bul. + Mac.	78.8	75.0	73.3	72.5	73.0
SC. + Sln.		80.8	76.8	77.0	73.7
Czsl.			86.8	83.0	75.7
Lus.				81.5	75.2
Pol.					77.7

It is apparent that the West Slavic languages form a branch consisting of Polish and the compact unit of Lusatian and Czech-Slovak, considering the high score 86.75% between latter subgroups. Slovenian is in a special position between Serbo-Croatian (85%) and Czech (84%). Naturally, it is not possible to separate Czech and Slovak. That is why it is necessary to evaluate the Czech-Slovenian relation from the Czech-Slovak perspective. The average of Czech-Slovak vs. Slovenian scores is 82%, and it is less than 85% for Slovenian vs. Serbo-Croatian on the one hand, still less than the average for all 5 West Slavic languages (86.2%), and even less than the average of the lowest scores within West Slavic, Polish vs. Lusatian and Polish vs. Czech-Slovak, namely (83.0+81.5)%/2 = 82.3%. And so it is necessary to accept the traditional affiliation of Slovenian together with Serbo-Croatian, although the position of Slovenian is more or less transitional. Interesting are the almost equal common proportions of cognates between West Slavic & Slovenian-Serbo-Croatian (78.4%) and Slovenian-Serbo-Croatian & Bulgar-Macedonian (78.8%), indicating a common Southwest Slavic dialect continuum, although the result 73.8% for the West Slavic branch and Bulgar-Macedonian is lower than the average score 75.9% for West and East Slavic and very close to 73.1% between South and East Slavic. This lowest result and the common arithmetic average 74.6% between East and Southwest Slavic define the period of the disintegration for all Slavic languages. Čejka's results may be depicted by the following tree-diagram (Čejka did not present any diagram of this type, but his data became a source for the diagram created by Girdenis, Mažiulis 1994, 11; the model of divergence presented here is based on the preceding discussion):

Diagram 1



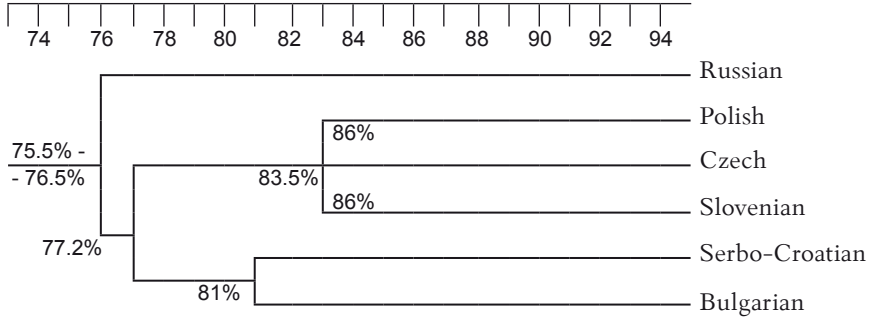
3.1.2. Another scholar who tried to apply 'classical glottochronology' to the Slavic languages, was the German J. Vollmer. His results were published by Johann Tischler in his monograph *Glottochronologie und Lexikostatistik* (Innsbruck 1973, 133). Vollmer compared 6 modern Slavic languages, plus Old Church Slavonic (his word-lists were not published):

Table 7

	Bul.	SC.	Slk.	Cz.	Pol.	Rus.
OCSl.	75	81	80	81	78	80
Bul.		81	81	74	72	74
SC.			82	77	77	77
Slk.				86	81	79
Cz.					86	76
Pol.						74

Abstracting from Old Church Slavonic as an extinct literary language, Vollmer's results can be depicted as follows:

Diagram 2



It is apparent that the topology of the diagram based on Vollmer's data is in principle in good agreement with Čejka' results, perhaps only the equality of Czech-Slovak and Czech-Polish is rather surprising. But both models, translated into the absolute chronology according to Swadesh's scenario, give, too young and thus ahistorical results: Čejka (74±1)%, i.e. AD 1000, Vollmer (75±0.5)%, i.e. AD 1050 as the date of disintegration of the Slavic languages.

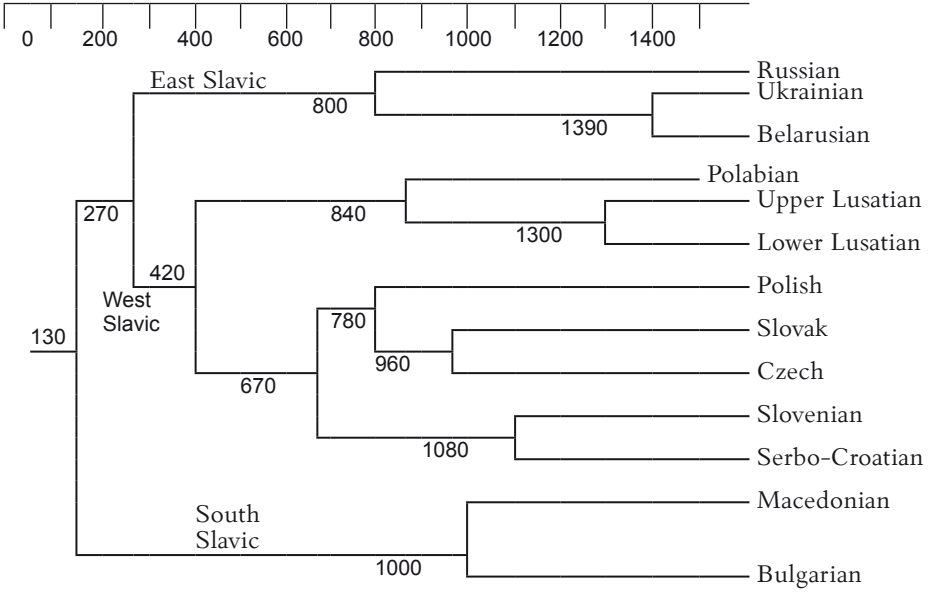
3.2. Let us compare the results based on 'classical glottochronology' with the results reached by applying the recalibrated glottochronology:

3.2.1. The first model was developed directly by Sergei Starostin with his team. We are grateful him for unpublished data from his database.

Table 8

	Mac.	SC.	Sln.	Slk.	Cz.	ULus.	LLus.	Plb.	Pol.	Blr.	Ukr.	Rus.
Bul.	90	88	84	82	81	75	75	77	80	82	76	80
Mac.		90	83	79	82	79	79	83	81	84	78	81
SC.			93	89	89	83	82	88	86	88	82	84
Sln.				87	90	82	81	88	86	85	79	85
Slk.					91	85	87	85	90	91	85	83
Cz.						89	88	88	88	87	80	82
ULus.							96	89	85	86	78	80
ULus.								90	89	86	79	80
Plb.									87	86	81	83
Pol.										90	85	85
Blr.											97	92
Ukr.												88

Diagram 3. **Classification of the Slavic languages after S. Starostin (presented in Santa Fe, NM, USA, March 2004)**



The present tree-diagram was generated by a computer program prepared by Sergei Starostin in the late 1980s. A preliminary version of this model was published in Starostin's article *Methodology of Long-Range Comparison*, which was first published in the volume: V. Shevoroshkin (ed.) *Nostratic, Dene-Caucasian, Austric and Amerind*, Bochum 1992, 78, and later reproduced in the volume: V. Shevoroshkin, P.J. Sidwell (eds.) *Historical Linguistics & Lexicostatistics*, Melbourne 1999, 65. The first version of the diagram still operated with the trichotomy, opposing East, West and South branches, but latter without Slovenian and Serbo-Croatian, which were classified together with the West branch.

3.2.2. The second model based on the 'recalibrated glottochronology' was prepared by the authors of the present study (Novotná 2004; Novotná, Blažek 2005). The word-lists cover 15 modern idioms, plus Polabian and Old Church Slavonic. In contrary to Starostin our calculation was realized 'manually', not via any computer program, but in agreement with the rules formulated by Starostin. The only methodological difference from Starostin consists in the systematic inclusion of synonyms. Swadesh postulated choosing only so called 'main' synonyms, the most frequent equivalents of concrete semantic units. But if there are more synonyms and some of them are related, the degree of the mutual genetic relationship is higher. And so it is not correct to eliminate synonyms. That is why we operate with 100 semantic units, while the number of the lexical units is usually higher. From our personal communication we know that Starostin also operated with synonyms, but not systemically. He also did not explain how to calculate with them. Our strategy is based on the standard list of 100 semantic units chosen already by Swadesh in 1955. The number of semantically identical and unborrowed units, attested in both compared languages, i.e. N_0, corresponds to 100%. The numerator in our proportion is represented by the number of all cognates, including synonyms.

Our results are summarized in table 9:

Table 9

	Bul.	Mac.	Srb.	Cr.	Sln.	Slk.	Cz.	ULus.	LLus.	Plb.	Kaš.	Pol.	Blr.	Ukr.	Rus.
OCSl.	90/ 100 0.900	88/ 100 0.880	90/ 99 0.909	92/ 100 0.920	90.5 100 0.905	94/ 99 0.949	96/ 99 0.970	92/ 99 0.909	90/ 99 0.909	77/ 88 0.875	85/ 97 0.876	88/ 99 0.889	81/ 97 0.835	81/ 99 0.818	85/ 100 0.850
Bul.		96/ 100 0.960	91.5/ 99 0.924	92.5/ 100 0.925	89/ 100 0.890	85/ 99 0.859	85/ 99 0.859	86/ 99 0.869	85/ 99 0.859	70/ 88 0.795	81/ 97 0.835	83/ 99 0.838	81/ 97 0.835	79/ 99 0.798	83/ 100 0.830
Mac.			91/ 99 0.919	92/ 100 0.920	88.5/ 100 0.885	83/ 99 0.838	83/ 99 0.838	84/ 99 0.848	84/ 99 0.848	71/ 88 0.807	78/ 97 0.804	81/ 99 0.818	81/ 97 0.835	79/ 99 0.798	83/ 100 0.830
Srb.				99/ 99 1.000	95.5/ 99 0.965	86.5/ 98 0.883	86.5/ 98 0.883	86.5/ 98 0.883	87/ 98 0.888	74/ 87 0.851	82.5/ 96 0.859	82.5/ 98 0.842	82/ 96 0.854	80/ 98 0.816	85.5/ 99 0.864
Cr.					98.5/ 100 0.985	89.5/ 99 0.904	90.5/ 99 0.914	89.5/ 99 0.904	90/ 99 0.909	78/ 88 0.886	84.5/ 97 0.871	85.5/ 99 0.864	85/ 97 0.876	83/ 99 0.838	88.5/ 100 0.885
Sln.						88/ 99 0.889	90/ 99 0.909	88/ 99 0.889	88.5/ 99 0.894	76.5/ 88 0.869	84.9/ 97 0.866	86/ 99 0.869	83.5/ 97 0.861	81.5/ 99 0.823	86/ 100 0.870
Slk.							97/ 99 0.980	93/ 99 0.939	92/ 99 0.929	74/ 88 0.841	85/ 96 0.885	89.5/ 98 0.913	85/ 96 0.885	84/ 98 0.857	86/ 99 0.869
Cz.								92/ 99 0.929	91/ 99 0.919	76/ 88 0.864	84/ 96 0.875	89.5/ 98 0.913	85/ 96 0.885	83/ 98 0.847	86/ 99 0.869
ULus.									98/ 99 0.990	77/ 88 0.875	86/ 96 0.896	89.5/ 98 0.913	86/ 96 0.896	84/ 98 0.857	88/ 99 0889
LLus.										77/ 88 0.875	86/ 96 0.896	91.5/ 98 0.934	85/ 96 0.885	83/ 98 0.847	87/ 99 0.879
Plb.											76/ 87 0.874	74/ 87 0.851	70/ 86 0.814	71/ 88 0.807	74/ 88 0.841
Kaš.												96/ 96 1.000	80/ 94 0.851	80/ 96 0.833	82/ 97 0.845
Pol.													84/ 96 0.875	84/ 98 0.857	84/ 99 0.848
Blr.														96/ 97 0.990	93/ 97 0.959
Ukr.															91/ 99 0.919

In the following steps we will abstract from Old Church Slavonic as an old literary (and rather artificial) language with an incomplete lexical corpus (the same may be said about Polabian; for this reason its results are rather problematic). The unexpectable share 93.2% connecting Old Church Slavonic with Czech requires a special explanation which is not a subject of the present study. Let us order the languages in groups, usually in pairs, according to languages with the closest relationship: Srb.-Cr. (= SC.) and Kaš.-Pol. agree 100%; regarding the different distribution of synonyms, they will be taken into account separately. Further ULus.-LLus. (= Lus.) 99%, Blr.-Ukr. 99%, SC.-Sln. 98%,.Cz.-Slk. 97%, Bul.-Mac. 95%. The comparison of Russian vs. Belarusian & Ukrainian gives 92.9%, indicating the East Slavic (= ESl.) unit.

The results of the comparison between these groups are summarized in table 10.

Table 10

	SC.-Sln.	Cz.-Slk.	Lus.	Plb.	Kaš.-Pol.	ESl.
Bul.-Mak.	92.0	86.9	86.9	80.7	84.2	82.8
SC.-Sln.		90.4	89.2	86.0	86.9	83.3
Cz.-Slk.			91.4	85.4	90.0	85.3
Lus.				88.0	92.5	86.4
Plb.					85.6	82.3
Kaš.-Pol.						85.2

The East Slavic unit was already defined. It is apparent that the South Slavic unit with the average score 92.0% in the BTL exists too. It is more than 89.2% between SC.-Sln. and Cz.-Slk. For the existence of the West Slavic (= WSl.) unit there are also the arguments: 91.3% without Polabian, 89.6% including Polabian. The final step is the comparison of the South, West and East branches of Slavic, in table 11a without Polabian, in table 11b with Polabian:

Table 11a

	WSl.	ESl.
SSl.	87.4	83.1
WSl.		85.7

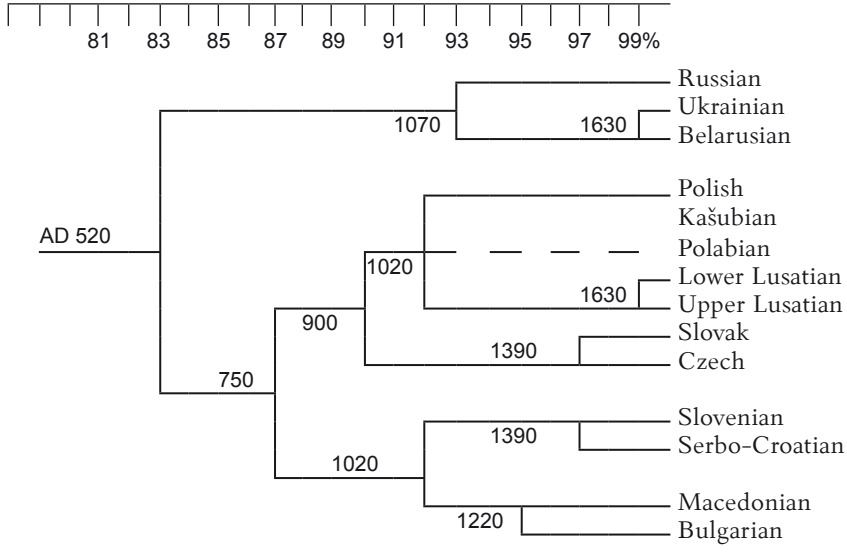
Table 11b

	WSl	ESl.
SSl.	87.0	83.1
WSl.		85.2

This means that the traditional trichotomic classification of the Slavic languages should be corrected. In contrary to the usual three equidistant units it is necessary to introduce a hierarchic model with a sequention of

two dichotomies. The first division separated the ancestors of the East and Southwest Slavic dialects, the second division separated West and South Slavic. The average of all scores gives the result 85.7% without Polabian and 85.5% with Polabian. The dating of the disintegration of the Slavic dialect continuum should be defined by the value of the lowest result 83.1%, reached for South and East Slavic. Translated into absolute chronology (see table 4 calculated by Starostin), it is possible to date the disintegration of the Slavic languages to AD 520. The West and South Slavic languages were separated in the middle of the 8th cent., West Slavic began its disintegration in the end of the 9th cent. and during 10th cent., South Slavic in the beginning of the 11th cent. and East Slavic around 1070. The position of Polabian is between Lusatian (88.0%), Czech (87.8%) and Polish-Kašubian (85.6%). Remarkable is the low score between Polabian and Slovak (83.0%) in comparison with Czech, and the high score between Polabian and Slovenian-Serbo-Croatian (86.0%). The mutual relations are depicted in diagram 4:

Diagram 4



The chronology of the following divergencies is difficult, regarding the phenomenon of 'dialect' chain. This chain appears, if we order the closest idioms in the direct neighbourhood:

LLus. Plb. Ukr.
99| |88.5 |99
Bul.-95-Mac.-94-Cr.-98-Sln.-92-Cz.-92-ULus.-93-Pol.-88.5-Blr.-94-Rus.
|97
Slk.

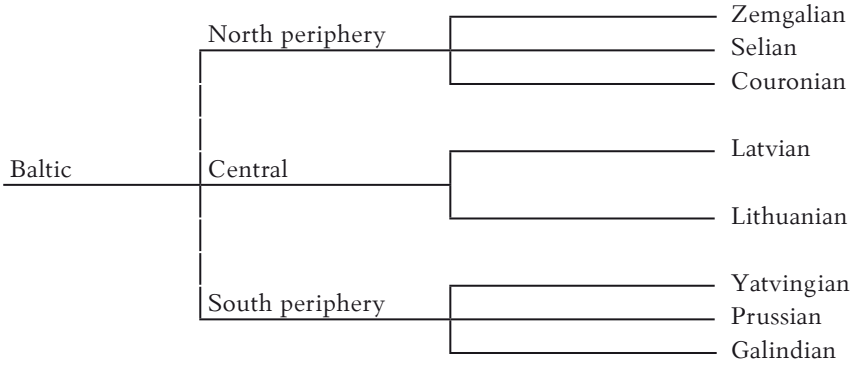
The scheme is more linear, if the common units Serbo-Croatian, Czech-Slovak, Lusatian and Belarusian-Ukrainian are taken in account (Polabian was left aside for its incomplete lexicon).

Bul.-95-Mac.-93-SC.-97-Sln.-91-Cz.+Slk.-91.5-Lus.-92.5-Pol.+Kaš.-86-Blr.+Ukr.-93-Rus.

Only in two cases do the figures fall under 90%. It is symptomatic that the lowest values indicate the limits between the south and west branches (91%) and west and east branches (86%). This means that this alternative approach gives the same results as the preceding steps, i.e. the divergence of the Slavic languages can be described as a sequence of two dichotomies: (1) east vs. southwest (6th cent.); (2) south vs. west (middle of the 8th cent.).

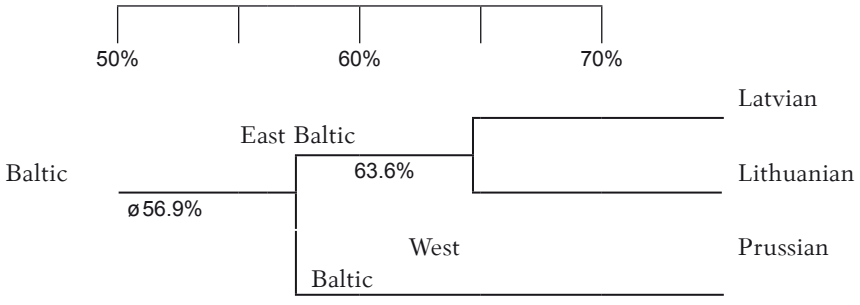
4. According to tradition, the Baltic languages are divided into a western part represented by Old Prussian, extinct from c. 1700, and an eastern part, represented by the living languages, Lithuanian and Latvian. But Baltic dialectology was much more complex a millennium ago. The following model was proposed by V. Mažiulis (1981):

Diagram 5



4.1. The first serious application of lexicostatistics (with 140-word-list, reduced for the limited Prussian lexicon) was used by Lanszweert (1984, xxxii–xxxvii), who found 63.6% for Lithuanian vs. Prussian, 58,6% for Prussian vs. Lithuanian and 55,2% for Prussian vs. Latvian:

Diagram 6



4.2. The results of Girdenis, Mažiulis (1994, 9) are rather different:

Table 12

	Latvian	Prussian
Lithuanian	68	53.6 /49.0*
Latvian		44.3

Note: The figure 49.0% is a result of the correction 0.490 = 0.536 · 0.915, where the latter coefficient expresses the age 600 years of most of the Prussian records.

The study of Girdenis & Mažiulis is also valuable for the individual comparison of Lithuanian, Latvian and Prussian with 12 Slavic languages:

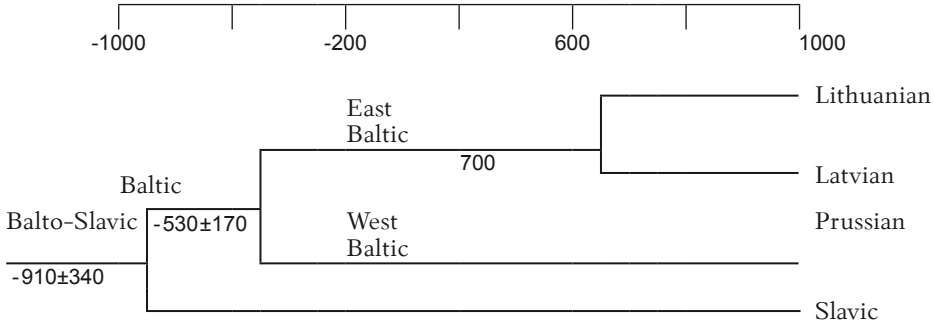
Table 13

	Bul.	Mac.	SC.	Sln.	Slk.	Cz.	ULus.	LLus.	Pol.	Blr.	Ukr.	Rus.
Li.	46	45	44	44	46	44	45	46	43	47	47	47
La.	42	41	41	40	42	41	45	43	40	44	40	45
Pr.	49!	39	41	40	42	42	42	42	39	40	41	41

Note: The figure 49% between Bulgarian and Prussian is apparently mistaken, probably it has to be 39%

Using their own data for the Baltic languages and Čejka's data for the Slavic languages and applying 'classical glottochronology', Girdenis, Mažiulis 1994, 11 proposed the scheme:

Diagram 7



4.3. Starostin (Workshop "Quantitative methods in Classification of Languages and Human Populations"; Santa Fe, NM, 2004, and p.c., June 2005) dated the separation of Lithuanian and Latvian to 80 B.C., Lithuanian and the 'Dialect of Narew' to 30 B.C., Latvian and the 'Dialect of Narew' to 230 B.C. The position of Prussian in his calculations is rather strange, it has to be closer to Slavic than to Baltic. The disintegration of the Balto-Slavic unity was dated to 1210 BC.

4.4. Our results were reached on the basis of the lexical data, compiled in the Appendix 1. Table 14 summarizes the mutual scores between the Baltic languages, table 15 between the Baltic and Slavic languages:

Table 14

language / %	Latvian	Prussian	'Narewian'
Lithuanian	84.8	62.0	76.5
Latvian		55.2	76.1
Prussian			43.0

Table 15

%	Bul.	Mac.	Srb.	Cr.	Sln.	Slk.	Cz.	ULus.	LLus.	Plb.	Kaš.	Pol.	Blr.	Ukr.	Rus.
Li.	49.0	48.0	48.5	49.0	48.0	51.5	51.5	50.5	48.5	47.7	48.5	49.5	50.5	49.5	50.0
La.	43.4	43.4	43.9	44.4	45.4	44.9	45.9	44.9	42.8	43.7	43.8	43.9	43.8	42.9	43.4
Pr.	49.4	48.3	49.9	49.4	48.3	50.4	52.5	50.4	48.3	47.4	48.9	48.9	46.7	46.7	46.2
Nar.	44.0	44.0	44.9	45.9	48.8	45.0	46.6	44.7	43.1	43.0	48.8	45.9	42.1	42.1	42.1

Table 16 demonstrates the average scores between South, West, East & all Slavic and the individual and all Baltic languages:

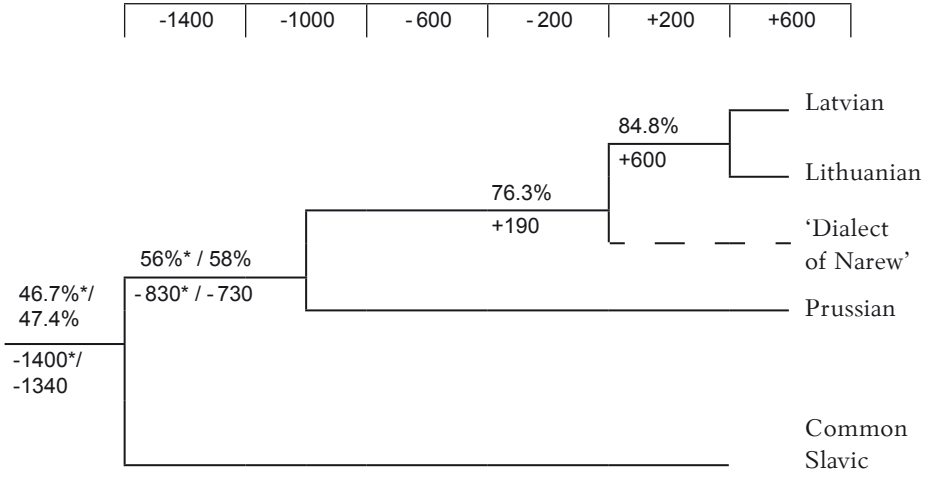
Table 16

	South Slavic	West Slavic	East Slavic	all Slavic
Lithuanian	48.5	49.7	50.0	49.4
Latvian	44.1	44.3	43.4	44.1
Prussian	49.0	49.5	46.5	48.7
'Narewian'	45.5	45.3	42.1	44.7
all Baltic	46.8 / 47.2*	47.2 / 47.8*	45.5 / 46.6*	46.7 / 47.4*

Note: *Without 'Narewian.'

Applying the 'recalibrated glottochronology' and including a calculation of synonyms, we reach diagram 8:

Diagram 8



4.4.1. The double result 58/56% for Prussian vs. the other Baltic languages reflects the calculation without / with the 'Dialect of Narew' (*Pogańske gwary z Narewu*; see Zinkevičius 1984). The score 43% between Prussian and the 'Dialect of Narew' in comparison with 62% and 55.2% for Prussian vs. Lithuanian and Prussian vs. Latvian respectively, excludes the identification of the 'Dialect of Narew' with the historical Yatwingians, known from the Middle Ages, if their language is to be connected with the other Baltic idioms of the southern periphery, including Prussian. Regarding this big difference, it seems better to accept the explanation

of Schmid (1986) who identified in the 'Dialect of Narew' a strong influence of Northeast Yiddish, spoken in the big cities of Lithuania and Latvia, hence the hybrid East Baltic-German idiom. For the relatively big difference between the Prussian-Lithuanian and Prussian-Latvian scores, viz. 62.0% vs. 55.2% respectively, there are at least two explanations: (i) The mutual influence between Prussian and Lithuanian, caused by their geographical proximity. (ii) The areal influence of Balto-Fennic or East Slavic on Latvian. In the analyzed 100-word-list, there is only one apparent borrowing of East Slavic origin in Latvian, viz. *cilvēks* "person, human being" and nothing from Balto-Fennic. This one item plays a minimum role. That is why it is necessary to admit a stronger role of mutual influence between Prussian and Lithuanian. For this reason, the separation of the central dialect, the ancestor of Lithuanian & Latvian, and the southern dialect, the ancestor of Prussian, should be closer to the result indicated by the score between Prussian & Latvian, i.e. 55.2%, reflecting 920 BC as the date of divergence with correction for the age of the Prussian language fragments (the coefficient 0.985 corresponding to the date *c.* 1400).

5. We have compared four attempts to apply glottochronology for the Slavic languages. All agree in the conclusion that the most divergent groups are East Slavic and Bulgarian-Macedonian. In three cases East Slavic is identified as the first separated branch, only Starostin saw Bulgar-Macedonian in this role. Applying 'classical glottochronology', Čejka and Vollmer reached very young data of divergence of Common Slavic – c. AD1000 (similarly Fodor – it was in fact his main objection against the method). Starostin's dating to AD130 represents the opposite extreme. Without any reference in the historical documents it is necessary to use indirect evidence to verify it. The counter-argument may be sought in the stratum of archaic Germanic borrowings in Common Slavic, which have been ascribed to the Goths (cf. Kiparsky 1934, 192f). The most intensive contact was probably realized from the middle of the 4th cent., when the Slavs were integrated into the tribe union, formed by the Gothic king Ermanaric, as described by the Gothic historian Jordanes writing in the middle of the 6th cent. (*Get.* §119: *Post Herulorum cede item*

Hermanaricus in Venethos arma commovit, qui, quamvis armis despecti, sed numerositate pollentes, primum resistere conabantur. Sed nihil valet multitudo inbellium, praesertim ubi et deus permittit et multitudo armata advenerit. Nam hi, ut in initio expositionis vel catalogo gentium dicere coepimus, ab una stirpe exorti, tria nunc nomina ediderunt, id est Venethi, Antes, Sclaveni; qui quamvis nunc, ita facientibus peccatis nostris, ubique deseviunt, tamen tunc omnes Hermanarici imperiis servierunt). Elsewhere Jordanes informs us about the Slavic settlement of the first half of the 6th cent.: *Introrsus illis Dacia est, ad coronae speciem arduis Alpibus emunita, iuxta quorum sinistrum latus, qui in aquilone vergit, ab ortu Vistulae fluminis per immensa spatia Venetharum natio populosa considet. Quorum nomina licet per varias familias et loca mutentur, principaliter tamen Sclaveni et Antes nominantur. Sclaveni a civitate Novitunense et lacu qui appellatur Mursiano usque ad Danastrum et in boream Viscla tenus commorantur: hi paludes silvasque pro civitatibus habent. Antes vero, qui sunt eorum fortissimi, qua Ponticum mare curvatur, a Danastro extenduntur usque ad Danaprum, quae flumina multis mansionibus ad invicem absunt* (*Get.* §§34–35). From both passages it is apparent, that Jordanes recognized three ethnonyms relating to the Slavs: *Venethi, Antes, Sclaveni*. They cannot all reflect synonyms, since only *Antes* are localized between the rivers Dniestr and Dniepr. The *Venethi* must have lived left (i.e. west?) of the northern branch of the Carpathian Mountains (*Alpes*) and the source of the Vistula river. And the territory inhabited by the *Sclaveni* was defined by the city *Novietunense*, the Mursian lake and the rivers Vistula/Viscla and Danaster, i.e. Dniestr [§35]. This means that the territory of the *Venethi* was a part of the territory of the *Sclaveni*, complementary to the *Antes*. It is almost generally accepted that the *Antes* represented the ancestors of the East Slavs (e.g. Niederle 1953, 145–47). It would imply the equation *Venethi* / *Sclaveni* = *non-Antes*. Briefly, the opposition *Antes* : *non-Antes* probably reflects the dichotomy East Slavic vs. Southwest Slavic. Jordanes' contemporary, the Byzantine historian Procopius of Caesarea in his work ΥΠΕΡ ΤΩΝ ΠΟΛΕΜΩΝ ΛΟΓΟΙ differentiated only Σκλαβηνοί and Ἄνται: Χρόνῳ δὲ ὕστερον Ἄνται καὶ Σκλαβηνοὶ διάφοροι ἀλλήλοις γενόμενοι ἐς χεῖρας ἦλθον, ἔνθα δὴ τοῖς Ἄνταις ἡσσηθῆναι τῶν ἐναντίων τετύχηκεν. But he was sure that they

still used the same language: ἔστι δὲ καὶ μία ἑκατέροις φωνὴ ἀτεχνῶς βάρβαρος (III, 14). The separation of the *Antes* = East Slavs can thus be interpreted as the result of the disintegration of the Common Slavic ethnic and dialect continuum.

5.2. The first archaeological culture, for which a direct development to the historical Slavs was proposed, is Trziniec-Komarov, localized from Silesia to Central Ukraine and dated to the period 1500–1200 BC (Gimbutas 1963, 61; Rybakov 1978, 182–96; Sedov 1979, 16; EIEC 338, 605–06; EIEC 526). This archaeological dating agrees with our glottochronological estimation of the disintegration of the Baltic and Slavic languages, c. 1400 BC. The separation of the ancestors of the Lithuanians & Latvians and Prussians, dated to the 9–8 cent. BC or better already to the 10 cent. BC (see above), correlates with the dating of the differences in the burial rites: after c. 1000 BC in the Southwest Baltic area the cremation was preferred, while in the East Baltic region inhumation burials continued (Kilian 1982, 47; EIEC 50). The reflex of the Slavic-Gothic symbiosis indicated by the stratum of East Germanic loanwords in Common Slavic, may be associated with at least one of the following cultures: *Przeworsk* from the territory of the upper Vistula-San-upper Dniestr, flourishing in the 2–4 cent. AD, *Zarubincy* from the basin of the upper Dniepr, dating from the 2 cent. BC to 2 cent. AD, *Černjaxovo*, known from the basins of the middle and lower Dniestr and Dniepr from the 2–5 cent. AD (EIEC 104–05, 470, 657; EIEC 526). The historically described Slavic expansion with its centre of gravity in the 6th cent. corresponds to the Prague & Penkov cultures. The Prague culture expanded in western Slavia (eastern Germany, Poland, Czech and Slovak Republics, Hungary, Romania, northwest Ukraine), the Penkov culture in eastern Slavia (in southern Ukraine, Moldova and Romania). The Penkov culture has been identified with *Antes* (EIEC 416, 448; EIEC 526).

6. Summing up, it is possible to reconstruct the prehistory and early history of the Balto-Slavic dialect continuum in time as follows:

15/14th cent. BC – crystalization of the proto-Slavs in the southern periphery of the proto-Baltic continuum, localized from Silesia to Central Ukraine (Trziniec-Komarov culture). Let us compare the

glottochronological estimates of the dates of divergence for some of the other Indo-European branches: Indo-Iranian – 2000 BC, Celtic – 1000 BC (Starostin; our date 1100 BC is very close), Germanic – 1st cent. BC, Tocharian – 1st cent. BC (see Appendix 2). These results represent unambiguous evidence for Balto-Slavic unity.

10/8th cent. BC – separation of the southwest Baltic dialect, the ancestor of Prussian, from the central Baltic dialect, the ancestor of Lithuanian and Latvian. The corresponding ancient communities differentiated in burial rites, namely the cremation vs. inhumation respectively.

200 AD – 5th cent. AD – coexistence of the Slavs and some East Germanic tribes (Goths?) in the territory from the upper Vistula and San to the middle Dniepr, i.e. including the probable Slavic homeland in the north and northeast of the Carpathian mountains.

6th cent. AD – Slavic expansion and first dialect differentiation between East Slavic (dialect of *Antes*) and the rest of Slavic. What was the first impuls for this disintegration? The migration and military activities of the Huns in Europe are probably too early (their power culminated in Europe in AD 375–453), on the other hand, the Avars came too late (568 is the date of their first conflict with the Byzantine Empire). Perhaps some of the East Germanic tribes, Goths or Gepids or both, occupying the territory between the Dniestr and the Carpathian Mountains, separated the *Antes* from other Slavs.

600 AD – separation of Latvian from the other central Baltic dialects, represented especially by Lithuanian. Regarding the phenomenon of Latvian palatalization, resembling the Slavic second palatalization, it is tempting to see here a specific Slavic influence, caused by the Slavic expansion, culminating in the 6th and 7th cent.

Note: So called *Pogańske gwary z Narewu* probably represent a hybrid idiom based on the interference of Lithuanian & Latvian and Northeast Yiddish (S c h m i d 1986). From the point of view of Baltic dialectology, their identification with Yatwingian seems to be excluded.

(To be continued in Blt 42(3))

Petra NOVOTNÁ, Václav BLAŽEK
Department of Linguistics & Baltic Studies
Faculty of Arts of Masaryk University
A. Nováka 1
CZ-60200 Brno
Czech Republic
[*petano16@seznam.cz*], [*blazek@phil.muni.cz*]