Classification of Selkup dialects in the 18th-19th centuries and language proximity analysis programs on LingvoDoc

Normanskaja Ju.
Doctor of Philology, Professor, Principal Researcher,
Head of the Laboratory "Linguistic Platforms"
of Ivannikov's Institute for System Programming RAS
Institute of Linguistic RAS Moscow, Russia
ORCHID 0000-0002-2769-9187

Summary*: The article explores the dialect material of the Selkup language collected in the 20th century by A.P. Dulzon and his followers, as well as by G.F. Miller, P.S. Pallas and M.A. Castrén in the 18th century. From the perspective of dialect-differentiating isoglosses, the article demonstrates that numerous exceptions existed in the correspondences between southern and central Selkup dialects. When considering materials based on well-known isoglosses, it is challenging to identify systematic differences between southern and central dialects in the 18th century. Therefore, the article aims to determine when the distinctions between southern and central dialects emerged and how significant the differences were between northern and central-southern Selkup. To address this issue, five dictionaries by G.F. Miller from the 18th century were uploaded to the LingvoDoc platform (lingvodoc.ispras.ru). Subsequently, in-house programs were used for the analysis of grapho-phonetic isoglosses and basic vocabulary. The results indicate that in the 18th century, the northern Selkup dialect was already distinct from central-southern dialects both phonetically and lexically. It can be argued that two Selkup languages existed during that period. During this time, there were no systematic phonetic isoglosses between southern and central dialects, but lexically, they already differed from each other. It was previously established that all phonetic features of the Narym dialect that can be identified in the 21st century were already present in the book by Saint Makary (Nevsky) "Materials for acquaintanceship with the dialect of the Ostyak people of the Narym region" written in 1887. Consequently, it becomes clear that phonetic differences between southern and central dialects only formed in the early 19th century, despite the glottochronological analysis by S.A. Starostin suggesting their separation in the mid-2nd millennium BCE.*

Keywords: *Selkup dialects, graphics, phonetics, lexicostatistics, archival dictionaries, audio dictionary.*

As noted by E.A. Helimski, the accumulation and systematization of Selkup language data, primarily collected by A.P. Dulzon and his followers, have allowed for greater clarity and completeness in understanding the dialectal division of the languages present in the Selkup territory. Helimski observes that the analysis of the first dialectal dictionaries, created in the 18th century based on the main dialect-differentiating isoglosses, indicated that the classification during that period was not fundamentally different. Refer to the literature review for more details [Helimski 2000: 68-79].

However, computer-based comparative analysis of the complete manuscripts of the first dictionaries and modern Selkup audio dictionaries, recorded from the last language speakers, demonstrates that over the past 250 years, the classification of the supposed dialects has indeed changed.

This is demonstrated by examples from the 20th century, collected by A.P. Dulzon and his students and published by V.V. Bykonya, illustrating the reflexes of the main dialect-classifying isoglosses identified by E.A. Helimski:

1) Southern *-m, -n,　-ŋ* vs. Central *-p, -t, -k* vs. Northern *-p/-m (-0), -t/-n (-0), -k/-ŋ (-0)*:

Indeed, this classification feature is not always observed in Selkup dialects in the 20th century according to [Bykonya 2005].

Table 1. The first dialect-differentiating feature according to E.A. Helimski in the materials of the Selkup dialects based on the dictionary [Bykonya 2005]

Translation	Northern	Central	Southern
'God, Sky'		нуп /Ob. Ch., Tym./ ноб /Ob. Sh., Ch., Vas., Tym., Tur./	ном /Ket., Ob. S., SH/
'Human'	кум /Tur./	қоп /Ob. Ch., Vas./ қуп, қуп /Ob. Sh., Ch., Vas., Tym./ **ком ~ қуп /Ob. Ch./** **қум ~ қуп /Ob. Sh/**	**қуп /Ket./** кум, қум /Ob. S., Sh.,Ket./ қом /Ob. S/
'Blood'	кым /Tur./ қэм /El./ кам / Tur./	қап /Ob. Ch./	кӭм /Ket./ қэм /Ket./ кам, қам /Ob. S., SH/
'Animal, Bird'	сӯрӭм /El./ сурун /Tur./	сурун /Tym./ уроп /Tym./ хуроп ~ хурун /Ob. Ch., Vas., Tym./	сурам /Ket./ сурум /Ket./ сурӭм /Ket./

Exceptions are indicated in the table in bold font: in the southern dialects, an innovative reflex, characteristic of the central dialects: *қуп* /Ket./, whereas in the central dialects, on the contrary, there are double reflexes: *ком ~ қуп* /Ob. Ch/, *қум ~ қуп* /Ob. Sh/, demonstrating an archaic reflex which is characteristic of the northern dialects according to E.A. Helimski.

In the first dictionaries of the 18th-19th centuries, the number of exceptions (double forms) in the central dialects is even greater.

Table 2. The first dialect-differentiating feature according to E.A. Helimski in the materials of Selkup dictionaries of the 18th-19th centuries.

Translation	Northern	Central	Southern
'God, Sky'	Pal. номъ Mil. nob	Mil. Nar. **num**, nub Mil. Tym. nub, nob Makary Narym. нуп, **нун, ном**	Pal. Ket., номъ Mil. Ket. nom Grig. Nizhnechain. номъ, нунъ Mil. Tom. пит
'Human'	Pal. кубъ, Mil. kub, gub	Makary Narym. коп, куп, **кум** Mil. Nar. **kum**, gᵏüb Mil. Tym. gᵏub	Mil. Tom. киит Mil. Ket. кит Makary Obsk.кам Pal. Ket., куммъ Grig. Nizhnechain. комъ, гомъ, кум, гумъ

| 'Blood' | Pal. *кемъ* | Pal. Narym. *кебъ* | Makary Obsk.*кам*
Pal. Ket., *кемъ*
Grig. Nizhnechain. *камъ* |
| 'Animal, Bird' | Pal. *суру* | Pal. Narym. *хуръпъ* | Pal. Ket., *сурумъ*
Grig. Nizhnechain. *су́румъ* |

2) Southern -0- vs. Central -γ- (-w-) vs. Northern -ŋ-:

According to the dictionary [Bykonya 2005], by the end of the 20th century a significant number of exceptions of the reflexes of this isogloss are found in both southern and central dialects. They are also indicated in the table in bold font.

Table 3. The second dialect-differentiating feature according to E.A. Helimski
in the materials of Selkup dialects based on the dictionary [Bykonya 2005]

Translations	Northern	Cantral	Southern
'knife'		*naγӭ* /Vas./	*naӭ* /Ket., Vas./
'summer'	*таңал ўссам* /Tur./ *таңы* /Tur./ *таңэң* /El./	**таӭт** /Ob. Ch./ *таγыт* /Vas., Tym./	**таγӭ** /Ob. S/ **таγаны** /Ket./ **таγын** /Ket./ *таыт* /Ob. Sh.,Ket./ **таγыт** /Ob.,Ket./

It is interesting that in the 18th-19th centuries, there were perhaps about as many exceptions, mainly concerning southern dialects. In these dialects, both in the 18th century and in the 20th century, the reflexes that were characteristic of central dialects, according to E.A. Helimski, were represented.

Table 4. The second dialect-differentiating feature according to E.A. Helimski
in the materials of Selkup dictionaries of the 18th-19th centuries.

Translations	Northern	Central	Southern
'knife'	Mil. *pánga*	Mil. Narym., Tym. *págge* Pal. Narym. *пaвы*	Mil. Tom. *pa* Pal. Ket., *na* Mil. Ket. **págge** Grig. Nizhnechain. *na*
'summer'	Pal. **таγынъ**		Pal. **таγынъ**

3) Southern, Central -w- (sometimes with further contraction) vs. Northern -m-:

In the use of this isogloss in the materials of A.P. Dulzon's card index, published in the dictionary [Bykonya 2005], there are significantly fewer exceptions. Only twice is the preservation of the archaic *-m-*, characteristic only of northern dialects, recorded in the intervocalic position in the Tym (central) and Ket (southern) dialects.

Table 5. The third dialect-differentiating feature according to E.A. Helimski
in the materials of Selkup dialects based on the dictionary [Bykonya 2005].

Translations	Northern	Central	Southern
'mother'		*ау* /Tym./ *эвэ̄* /Tym./	*äвäм* /Ket./ *авэ̄* /Ket./ *ава* /Ob. SH/ *авва* /Ob. S., SH/ *эввы, эввэ̄* /Ket./
'larch'	*тӱ̄мы* /Tur./	*тöвэ̄л пō* /Vas./	*тӱу* /Ob. S/ *тӱвэ̄* /Ket./
'rabbit'	*нёма* /Tur./	**нёма** /Tym./ *нэва* /Vas./ *нева* /Ob. Ch., Vas./	*нева* /Ob. SH/ *нё* /Ob. S.,Ket./
'to eat'	*амырӄо* /Tur., El./, *амырыӄо* /Tur./	*авешпэ̄гу* /Vas./ *авэшпэ̄гу* /Ob. Ch., Vas./ *авыргу* /Vas./ *авӭргу* /Ob. Ch./ *аоргу* /Ob. Ch./ *аургу* /Ob. Ch./	*аурукугу* /Ob. S/ **амыргу** /Ket./ авэшпугу /Ob. SH/ авӭргу /Ket./

In the first dictionaries and books, the number of exceptions is also insignificant. Mostly, contracted forms are noted. However, it is significant that this isogloss does not serve as a differentiating factor between southern and central dialects, which tend to blend together, judging by the materials of the first two isoglosses.

Table 6. The third dialect-differentiating feature according to E.A. Helimski
in the materials of Selkup dictionaries of the 18th-19th centuries.

Translations	Northern	Central	Southern
'mother'	Kastr. *eme* 'mother' Pal. *амель* Mil. *hémed*	Makary Narym. **амба**, *авем* Pal. Narym. **амба** Mil. Narym. *éweb* Mil. Tym. *eweb*, **amma**	Makary, middle Ob. *аву*, *ау* Mil. Tom. ewe Pal. Ket., *евемъ* Mil. Ket. ewem Grig. Nizhnechain. *ау*
'larch'		Mil. Nar., тым *töo*	Mil. Tom. *tiù* Mil. Ket. *tiiu*
'rabbit'	Mil. *njäma*	Mil. Nar., Tym. *njoa*	Mil. Tom. *njo* Mil. Ket. *njoa*
'to eat'	Pal. *аморко* Mil. *ámerlang*	Pal. nar. *аурходъ*	Mil. Tom. *áurak* Pal. Ket., **агургу**

4) Southern *s*- vs. Northern Central *š*-

This feature, exemplified by the lexeme 'two', based on the materials of the dictionary [Bykonya 2005], has the greatest number of exceptions in the dialects of the second half of the 20th century. In northern dialects, not only *š-* but also *s-* can be found, and conversely, *š-* is found in central dialects and the southern Ket dialect. The reflex of *s-* from southern dialects is characteristic only for the Ob Sosyugum dialects according to [Bykonya 2005].

Table 7. The fourth dialect-differentiating feature according to E.A. Helimski
in the materials of Selkup dialects based on the dictionary [Bykonya 2005].

Translations	Northern	Central	Southern
'two'	*шиты* /Tur./ *шиты* /Tur./ **ситэ** /El./ **ситтэҡ** /El./ **сим** /Tur./	*шэдӭйя* /Ob. Ch./ *шӭдӭя* /Ob. Ch./ *шэдэгӭ, шэтэҕӭ* /Vas./ *шэдэҕ* /Ob. Ch./ *шӭдэҕ* /Ob. Ch./ *шöт* /Ob. Ch./ *шӭдӭ* /Ob. Ch., Тym./ *шӭдӭя* /Ob. Ch./ *шыдäҕ, шыдäях, шэдäҡ,* *шэдэäҡ* /Тym./ *шӭдэҡ* /Vas./ *шытэҕ* /Vas./ *шэдэäҡ* /Тym./ *шӥдия* /Тym./ *шыдäях* /Тym./ *шэдäҡ* /Тym./ *шэдэ* /Vas./ *шэтӭ* /Vas./ *шидӭ, шэдэ, шэтӭ* /Vas./ *шытэҕ* /Vas./	*сöдияҡ* /Ob. S/ *сыдӭяҡ* /Ob. S/ *сыдаҡ, сыдӭяҡ* /Ob. S/ *сыдӭ* /Ob. S/ *сыт* /Ob. S/ *сöд* /Ob. S/ *сэдӭ* /Ob. S/ **шэдӭйя** /Ob. SH/ **шэд** /Ob. SH/ **шиты** /Ket./ **шыдэк** /Ket./ **шиттӭ** /Ket./ **шыттэҡы** /Ket./ **шытäҡы** /Ket./ **шыттӭ** /Ket./ **шиты, шыттӭ** /Ket./ **шыдэк, шыттэҡы** /Ket./

In the first dictionaries (by G.F. Miller) and the book (by St. Makary), the reflex *š-,* which according to E.A. Helimski is characteristic only for the northern and central dialects, is also found in the southern Ket and Ob dialects.

Table 8. The fourth dialect-differentiating feature according to E.A. Helimski in the materials of Selkup
dictionaries of the 18th-19th centuries.

| Translations	Northern	Central	Southern
'two'	Mil. *schittjächan* Pal. *шитъ*	Mil. Nar., Тym. *schiddéga* Pal. nar. *шитыя*	Mil. Tom. *ssitáge* Makary Obsk.**шидо,** **шит, шиты,** **шиды-мчели**, Mil. Ket. **schiddéga** Pal. Ket., *ситыга*

Thus, this feature also does not reliably distinguish the three groups of Selkup dialects.

As a part of our work in the Russian State Archive of Ancient Documents in Moscow, we were able to digitize the originals of the earliest manuscripts of Selkup dictionaries written by G.F. Miller in the 18th century.

Currently, 5 handwritten dictionaries by G.F. Miller (Karasin, Tomsk, Ket, Tym, Narym), are interconnected with each other and with modern Selkup audio dictionaries through etymological connections. On LingvoDoc, using a special in-house program called "Analysis of cognates in different dialects of one language," which replicates the work of an etymologist, users who are registered in the system, and are either the creators of the dictionary or have received rights from the creator of the dictionary, can open the Tools option and select various options for analyzing the dictionary.

Figure 1: Using tools for dictionary analysis in LingvoDoc
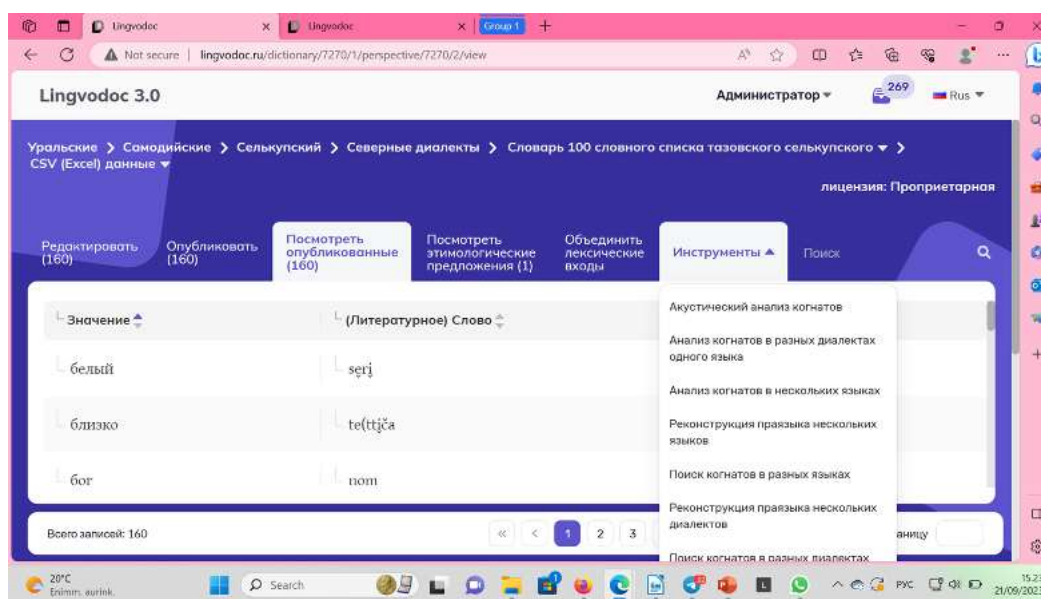
After selecting the option "Analysis of cognates in different dialects of one language" the user is presented with an interface allowing them to choose dictionaries that will be analyzed. See Figure 2.

Figure 2. Interface allowing the selection of which dialects cognates will be analyzed
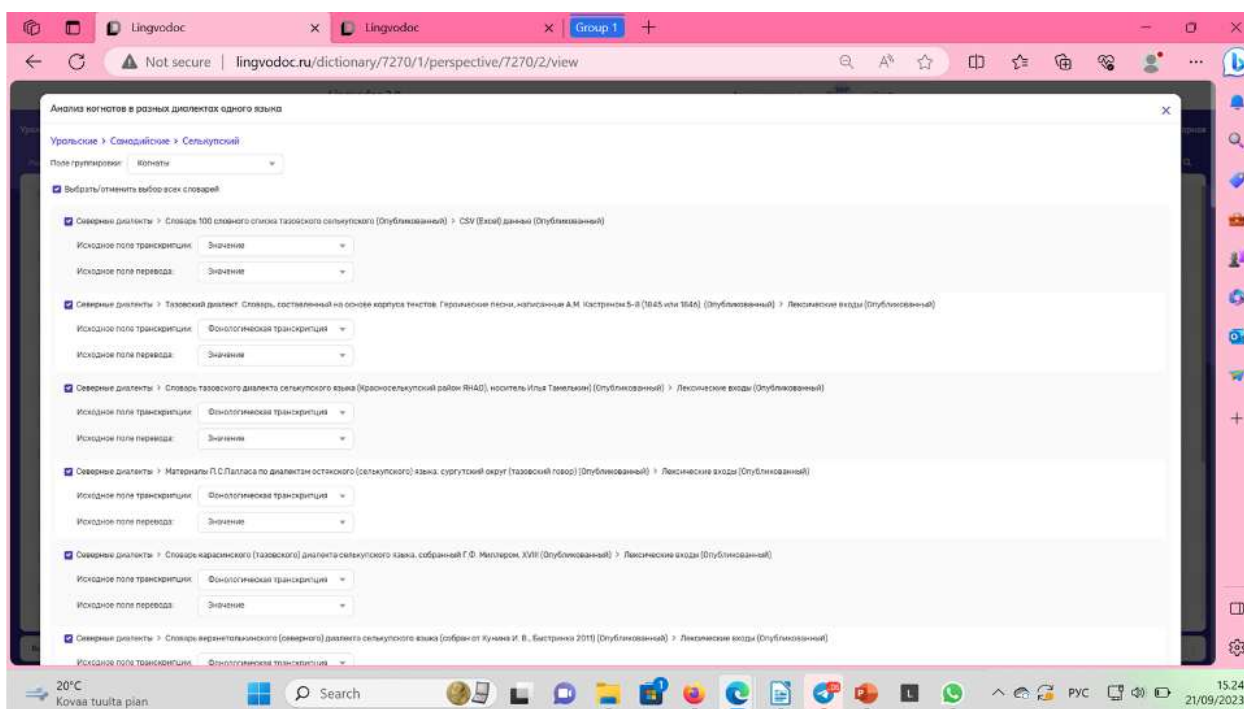
1) At the initial stage, the Cognate Analysis program computes correspondences for each symbol in the transcription algorithmically, comparing it with words from other dialects of the same language linked to the current dictionary:

   a) Roots are computed based on the principle that the first vowel (combination of vowels) corresponds to the first vowel (combination of vowels), the first consonant (combination of consonants) corresponds to the first consonant (combination of consonants), and the second consonant corresponds to the second one. As a result, a list of correspondences is generated for each pair of idioms. The dictionary author has the option to download it, analyze it, check the accuracy of transcriptions and etymologies that led to non-standard correspondence series, and make corrections to the transcription and etymology. Then, the algorithm is rerun on the revised material;

   b) The algorithm evaluates whether there are phonemes with two or more correspondences in the second dialect. If two or more phonemes from the considered dictionary correspond to one phoneme from another, the system checks whether there is a positional distribution between them that was not considered in the first stage. As a result of this calculation, the system outputs a list of correspondences between the two dialects in Excel format, with possible distribution rules.

2) Matching formant clouds of vowels with their correspondences:

   a) For each vowel with two or more correspondences in another dialect, the distribution of formant points representing vowels corresponding to the first vs. second vs. third, etc., phoneme on the graph is determined across the cloud. As a result, N 3D graphs are generated in Excel format (N = number of vowels), where points representing different vowels are marked with different colors on each graph. Each graph is accompanied by a decryption of words with translation, transcription, formant values, and an indication of the color of the point. If it turns out that on some graph for phoneme X, points of different colors intersect by less than 30%, the algorithm suggests splitting this phoneme into two or three different ones. The dictionary author has the option to download these graphs and algorithm suggestions for further study;

   b) Subsequently, dictionary authors have the opportunity to refine the transcription based on the data obtained. After making changes, the algorithm proceeds to the next stage.
Steps 1 and 2 can be repeated cyclically for all dialects of one language or for a series of related languages.

Thus, on LingvoDoc, grapho-phonetic isoglosses, distinguishing G.F. Miller's dictionaries (1151 lexemes from 5 dictionaries), were analyzed using the Cognate Analysis program.

Since 2023, on LingvoDoc it is possible to analyze dictionaries connected by etymological links, using lexical-statistical analysis programs. In the dictionaries' "Tools" tab, the option "Glottocronological analysis of languages/dialects" can be selected. This option can be applied to any set of dictionaries which contain   more than 50 words from M. Swadesh's word list. The 100-word Swadesh-Starostin list was selected, as only for this list strict semantic specifications have been justified and developed, allowing for sufficiently accurate comparative data analysis for different languages. According to S.A. Starostin's proposal of glottocronology, cf. [Starostin 1989], loanwords from other languages are removed, related words are linked on the LingvoDoc platform by etymological connections, then the percentage of matches between the lists of two idioms is calculated using the formula presented in Figure 3, and the approximate time of divergence is determined.

Figure 3. S.A. Starostin's formula for calculating
the proximity of languages and dialects, integrated into LingvoDoc

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

This formula, for which S.A. Starostin empirically selected lambda equal to 0.05, allows to determine the time of divergence for any set of languages, cf. [Starostin 1989]. LingvoDoc also has a built-in function for creating language similarity graphs in 2D and 3D formats.

Below are the results of the analysis using two methods: "Cognate Analysis" and "Glottocronology" applied to dictionaries created in the 18th century by G.F. Miller and field materials collected in the late 20th and early 21st centuries.

Part I. Results of processing G.F. Miller's dictionaries using the Cognate Analysis program

As mentioned earlier, 1151 lexemes from 5 dictionaries were analyzed using the Cognate Analysis program, see Figure 4.
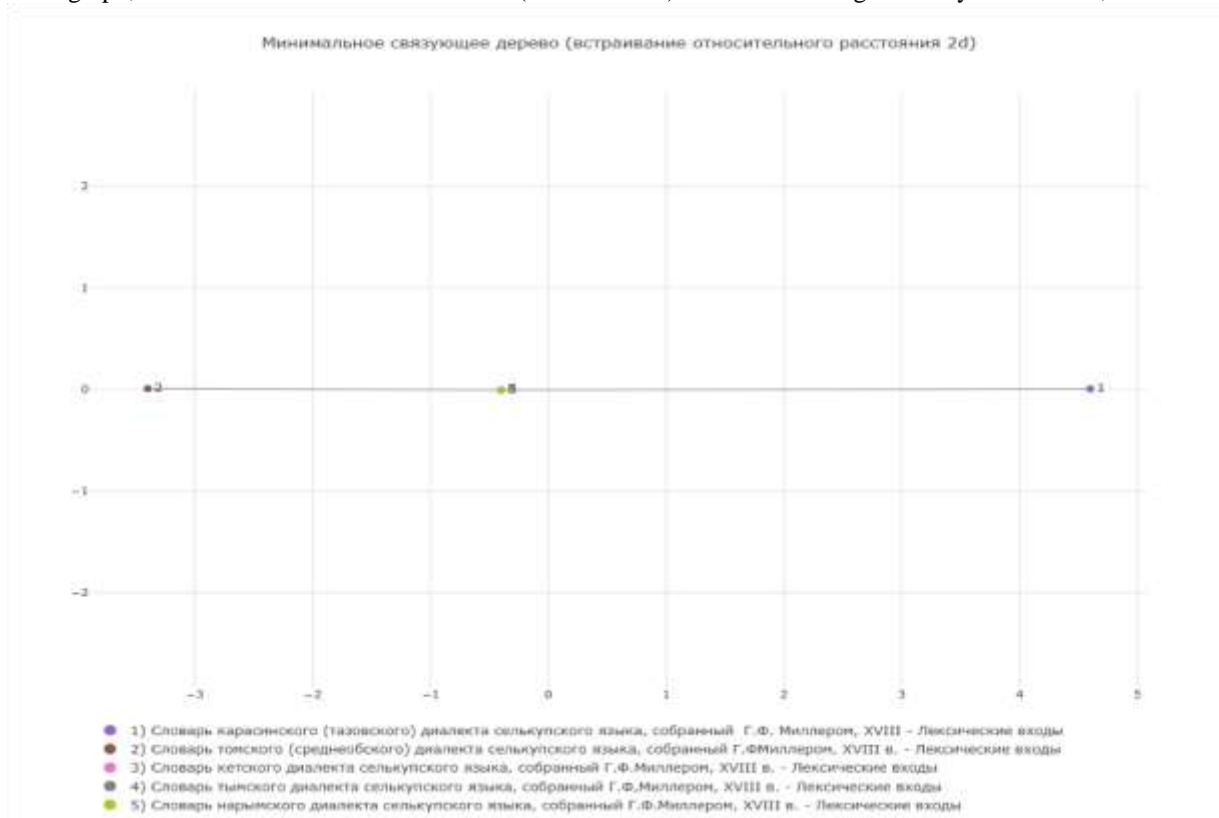
Figure 4. Number of analyzed lexemes from G.F. Miller's dictionaries
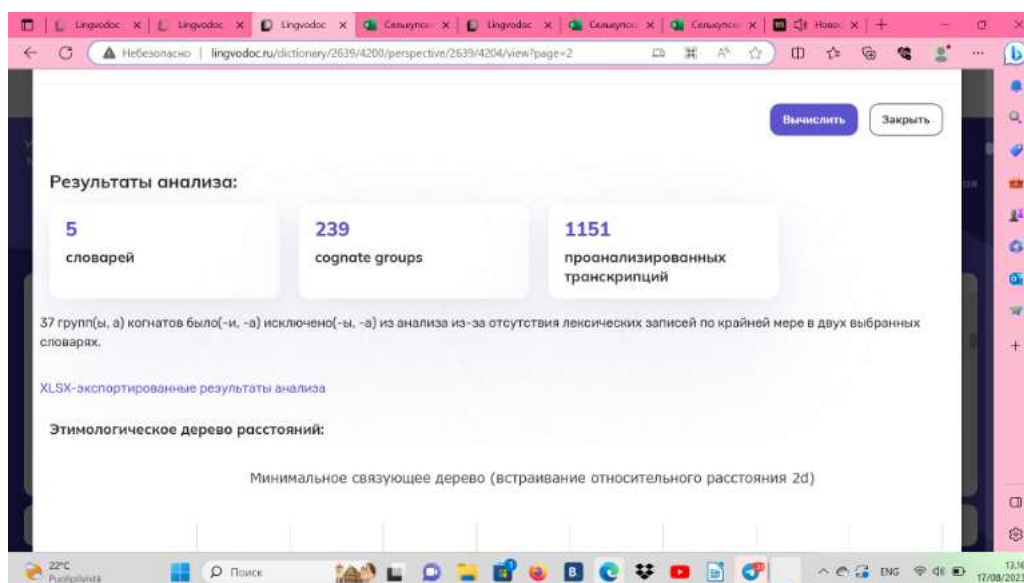
As a result of this analysis, the following cognate similarity graph was constructed, see Figure 5.

Figure 5. Cognate similarity graph of G.F. Miller's dictionaries

On the graph, it can be observed that the Karasin (Northern Taz) dialect differs significantly from others, while

the Ket (Southern), Tym, and Narym (Central) dialects essentially do not yet exhibit systematic dialect differentiating features, for which there would be more than three examples consistently displayed in all dialects. Certain differences are only observed in the now extinct Tomsk (Middle Ob) dialect. Specific isoglosses can be found in Table 9.

Table 9. Grapho-phonetic isoglosses distinguishing Selkup dialects in G.F. Miller's dictionaries

Karasin (Taz)	Ket	Tomsk (middle Ob)	Tym	Narym
Special reflexes in Karasin (Taz)				
h	*0*	*0*	*0*	*0*
-x-	*-g-*	*-g-*	*-g-*	*-g-*
i	*y*	*y*	*y*	*y*
ä	*a*	*a*	*a*	*a*
Special reflexes in Tomsk (middle Ob)				
ʧ-	*ʧ-*	***tj-***	*ʧ-*	*ʧ-*

-*tʃ*-	-*tʃ*-	**-d*ʃ*-**	-*tʃ*-	-*tʃ*-
t	t	**tj**	t	t
ʃ	ʃ/s	**sj**	ʃ	ʃ

On Figure 6, it can observed how the Cognate Analysis program provides examples illustrating Table 9,

Figure 6. Examples gathered in Selkup dialects, where "h-" is present in the Karasinsky dialect but absent in other dialects at the beginning of the word.

It is interesting to note that most regular differences identified in the 18th century differ from the commonly accepted isoglosses considered dialect-differentiating in the 20th-21st centuries.

Part II. Results of processing G.F. Miller's dictionaries using the Glottochronology program

As a result of the analysis of the 100-word lists from G.F. Miller's dictionaries, the following result was obtained, see Table 10.

Table 10. Results of glottochronological analysis of Selkup dictionaries by G.F. Miller.

	1. Dictionary of the Karasin (Taz) dialect of the Selkup language, collected by G.F. Miller, XVIII	2. Dictionary of the Tomsk (middle Ob) dialect of the Selkup language, collected by G.F. Miller, XVIII century.	3. Dictionary of the Ket dialect of the Selkup language, collected by G.F.Miller, XVIII century.	4. Dictionary of the Tym dialect of the Selkup language, collected by G.F. Miller, XVIII century.	5. Dictionary of the Narym dialect of the Selkup language, collected by G.F.Miller, XVIII century.
1	n/a	1.26 (86%)	1.35 (84%)	1.23 (86%)	1.26 (86%)
2	1.26 (86%)	n/a	1.26 (86%)	1.04 (90%)	0.94 (91%)
3	1.35 (84%)	1.26 (86%)	n/a	0.63 (96%)	0.65 (96%)
4	1.23 (86%)	1.04 (90%)	0.63 (96%)	n/a	-0.00 (100%)
5	1.26 (86%)	0.94 (91%)	0.65 (96%)	-0.00 (100%)	n/a

The proximity of G.F. Miller's Selkup dictionaries from the perspective of glottochronology can also be observed in the graph, see Figure 7.

Figure 7. Proximity graph of 18th-century Selkup dictionaries collected by G.F. Miller from the perspective of glottochronology.
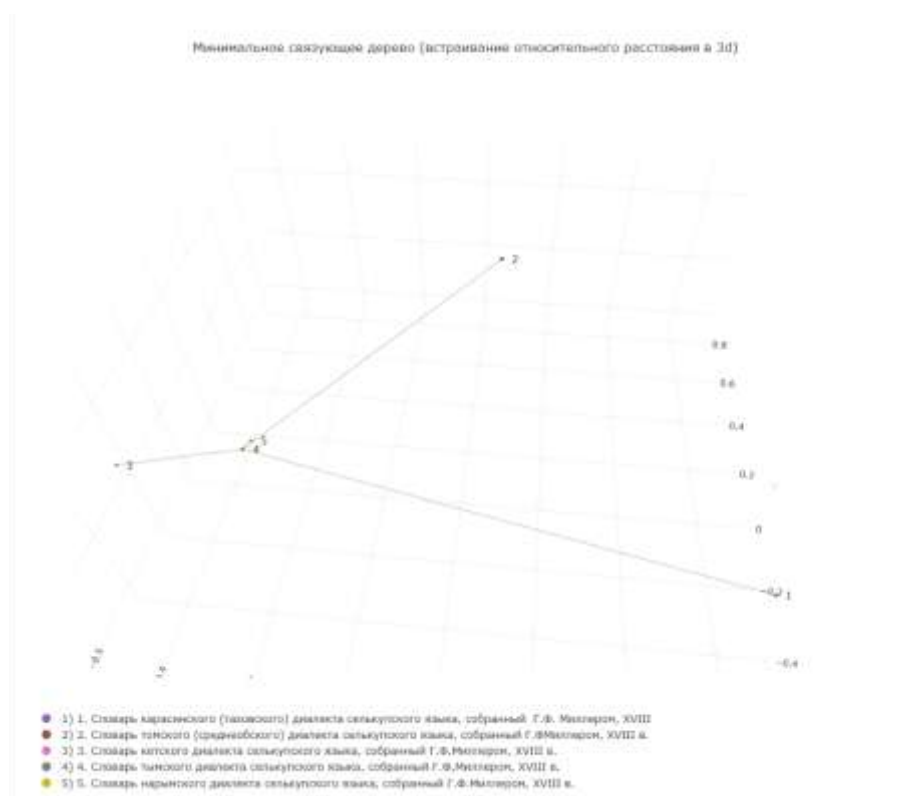
It is generally accepted, cf. [Burlak, Starostin 2005], that idioms with over 90% matches in basic vocabulary are considered dialects of the same language. We can observe that the Ket, Tym, and Narym dialects have approximately 96-100% matches, confirming their dialectal status, as indicated by the analysis of grapho-phonetic isoglosses. The Tomsk dialect also displays 90-91% matches with the Tym and Narym (central) dialects, and 86% similarity with the Ket (southern) dialect. The Karasin (northern Taz) dialect has less than 86% similarity with all dictionaries, suggesting that it was already a separate language in the 18th century. The dialects now classified as southern (Ket) and central (Tym and Narym) did not have systematic grapho-phonetic differences in the 18th century but already had slight differences in basic vocabulary.

List of abbreviations

Vas. — Vasyugan Selkup

Grig. Nizhnechain. — Nizhnechain selkup as found in the books by Grigorovsky N.P.

El. — Yelabuga Selkup

Kastr. — Taz Selkup as found in the books by M.A.Kastren

Ket. — Ket Selkup

Makary Narym— Narym Selkup as found in the book by St. Makary

Makary Obsk. — Obsk Selkup as found in the book by St. Makary

Mil. Nar. — Narym Selkup as found in the dictionary by G.F.Miller

Mil. Tom.   — Tomsk Selkup as found in the dictionary by G.F.Miller

Mil. Tym.   — Tym Selkup as found in the dictionary by G.F.Miller

Mil. Ket.   — Ket Selkup as found in the dictionary G.F.Miller

Ob. S. — the Ob dialects of Syusyukum

Ob. Ch. — the Ob dialects of Chumylkup

Ob. Sh. —   the Ob dialects of Sheshkup, Sheshkum

Pal. Ket. — Ket Selkup as found in the dictionary by P.S.Pallas

Pal. Narym. —   Narym Selkup as found in the dictionary by P.S.Pallas

Tur. — Turukhan dialect

Tym. — Tym dialect

Bibliography

[1].   Burlak, S.A., Starostin, S.A. (2005). Comparative-Historical Linguistics. Moscow.

[2].   Bykonya, V.V. (Ed.). (2005). Selkup-Russian Dialect Dictionary. Tomsk.

[3].   Helimski, E.A. (2000). Comparative Linguistics, Uralistics: Lectures and Articles. Moscow.

[4].   Kassian, A., Starostin, G., Dybo, A., Chernov, V. (2010). The Swadesh Wordlist: An Attempt at Semantic Specification. Journal of Language Relationship, 4, 46-89.

[5].   Starostin, G. (2010). Preliminary Lexicostatistics as a Basis for Language Classification: A New ApproaCh. Journal of Language Relationship, 3, 79-116.

[6].   Starostin, S.A. (1989). Comparative-Historical Linguistics and Lexicostatistics. Linguistic Reconstruction and Ancient History of the East, Part I. Moscow, 3-39.