DOI: 10.15514/ISPRAS-2025-37(3)-14

Glottochronology Classification of the Modern and the Earliest Samoyed Dictionaries using LingvoDoc Programs

Ju.V. Normanskaja, ORCID: 0000-0002-2769-9187 <julianor@mail.ru>

Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.
Institute of Linguistic RAS
1, B. Kislovskiy lane. Moscow 125009, Russia.

Abstract. The classification of Samoyedic languages has become one of the most popular topics in Uralistics in recent years, with at least six different perspectives expressed by leading experts, often in contradiction with one another. On the LingvoDoc platform, there are 16 dictionaries and concordances of texts in Samoyedic languages. Among these, 10 dictionaries – Nenets, Enets, Nganasan, and Selkup dialects – were compiled from native speakers, while six others were derived from archival and published sources. They are analyzed using the glottochronology formula developed by S.A. Starostin. The analysis on LingvoDoc results in a 3D graph that depicts the degree of temporal proximity regarding the divergence of Samoyedic languages and dialects. It was determined that, from a glottochronological perspective, there was a certain proximity between Nenets, Enets, and Nganasan, that are traditionally grouped into the North Samoyedic cluster, while Selkup, Mator, and Kamasin, are regarded as South Samoyedic. However, these commonalities existed for a relatively short period. A longer period of unity was observed between Mator and Kamasin languages and between Nenets and Enets. The highest number of words with no etymology in other lists of basic vocabulary was found in Selkup dialects and in the Nganasan language, indicating their prolonged isolated existence. The analysis conducted in this study supports the validity of the traditional classification of Samoyedic languages. Considering the material from early Selkup texts provides more reliable evidence for postulating the South Samoyedic group.

Keywords: Samoyed languages; Nenets; Enets; Nganasan; Selkup; Mator; Kamasin; classification; field research; archival data.

For citation: Normanskaja Ju.V. Glottochronology classification of the modern and the earliest Samoyed dictionaries using LingvoDoc programs. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025, pp. 195-210. DOI: 10.15514/ISPRAS-2025-37(3)-14.

Acknowledgements. The work was supported by grant RSCF No 25-78-20002**.**

Глоттохронологическая классификация современных и наиболее ранних словарей самодийских языков на платформе ЛингвоДок

Ю.В. Норманская, ORCID: 0000-0002-2769-9187 <julianor@mail.ru>

Институт системного программирования им. В.П. Иванникова РАН,
Россия, 109004, Москва, ул. Александра Солженицына, д. 25.
Институт языкознания РАН
Россия, 125009, Москва, Б. Кисловский пер., д. 1.

Аннотация. Классификация самодийских языков в последние годы является одной из наиболее спорных тем исследования. В последние десятилетия предложено как минимум 6 различных классификаций, которые противоречат друг другу. Нет ни одной дочерней группы самодийских языков, по поводу существования которой все авторы были бы согласны. В статье проанализированы как современные словари, собранные в экспедициях от последних носителей, так и наиболее ранние словники самодийских языков, доступные в глобальной сети посредством платформы ЛингвоДок.

Ключевые слова: самодийские языки; ненецкий; энецкий; нганасанский; селькупский; маторский; камасинский; классификация; полевые исследования; архивные данные.

Для цитирования: Норманская Ю.В. Глоттохронологическая классификация современных и наиболее ранних словарей самодийских языков на платформе ЛингвоДок. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025, pp. 195-210 (на английском языке). DOI: 10.15514/ISPRAS-2025-37(3)- 14.

Благодарности. Работа поддержана грантом РНФ № 25-78-20002.

1. Introduction

The classification of Samoyedic languages is one of the most popular topics in Uralistics in recent years, with at least six contradicting viewpoints expressed by leading specialists. In fact, there is not a single sub-group of Samoyedic languages on the existence of which all authors would agree.

Currently, the LingvoDoc platform [1] features a tool called the "Glottochronological Analysis of Languages", which analyzes wordlists of basic vocabulary using the methodology proposed by S.A. Starostin [2]. Online wordlists of modern Samoyedic languages based on archival materials from the 18th to the early 20th century, and recorded in audio form the last speakers of the languages, are available. It is believed that the glottochronological analysis of these sources, some of which were recorded from native speakers, will clarify the situation regarding conflicting classifications.

This work consists of two parts. The first part provides a brief analysis of existing classifications, while the second part presents the results of the analysis of 16 lists of basic vocabulary using S.A. Starostin's formula, which is integrated into the LingvoDoc platform.

2. Overview of Existing Classifications of Samoyedic Languages

The "Traditional" classification (see Fig. 1) was primarily based on the geographical location of the Samoyedic speakers, e.g., [3]. According to it, a Northern Samoyedic languages group is distinguished, which consists of: Nenets, Enets, Nganasan, and Southern Samoyedic languages: Selkup and Sayan Samoyedic.

This classification remains the most widely accepted to this day, see [4].

In the classification proposed by E.A. Helimski in [5], which is based on the analysis of a small list by M. Swadesh, consisting of 92 words of basic vocabulary in six Samoyedic languages, he concluded that the Southern Samoyedic languages did not form a group but separated independently during the divergence of the Proto-Samoyedic language.

In a 2004 report by S.A. Starostin (see [6]), wordlists collected by E.A. Helimski and supplemented in 2001 based on existing dictionaries were analyzed using a special formula developed by S.A. Starostin to refine the timing of divergence. As a result, he obtained results that essentially coincided with the traditional classification (see Fig. 1). The dating obtained by S.A. Starostin can be seen in Fig. 2 (the numbers in the figure indicate the date of the Proto language split, negative numbers indicate that the time of the splitting was BC). However, he came to the conclusion of the absence of Southern Samoyedic unity and the earliest divergence of the Selkup language.

In 1987, H. Katz expressed the opinion that Selkup and Kamasin should be grouped together, with Mator considered as the earliest separated Samoyedic language (see Fig. 3).

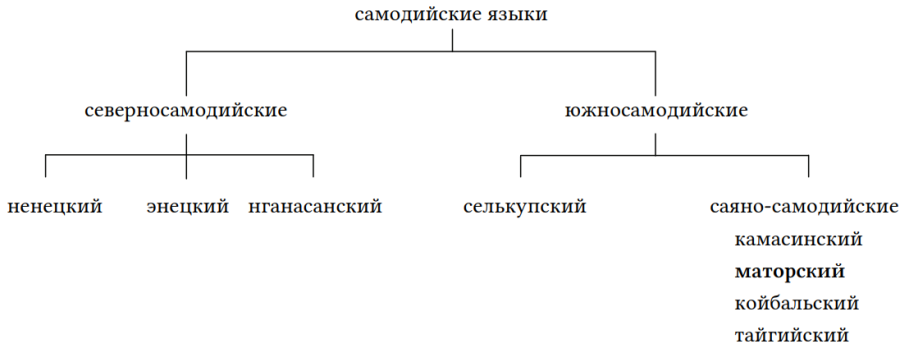

Fig. 1 "Traditional" classification of Samoyed languages.

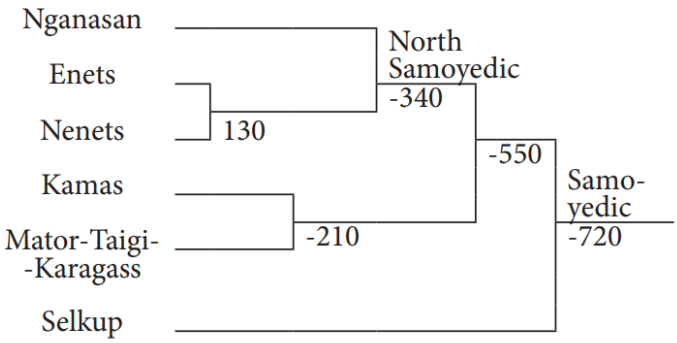

Fig. 2. Time of divergence of Samoyedic languages according to [6].

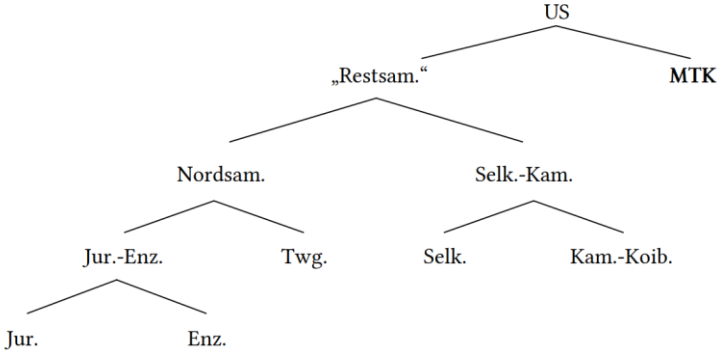

Fig. 3. Classification of Samoyedic languages according to [7].

In 1998, J. Janhunen proposed a different classification [8]. According to him, the Nganasan language separated first from the Proto-Samoyedic language, followed by Mator, and only then did the gradual divergence of other daughter languages occur, see [8] and Fig. 4.

J. Janhunen points out that this scheme corresponds well to the geographical locations of the language speakers, as the speakers of Nganasan and Mator languages lived in the northeastern and southeastern territories of the Samoyedic language speakers.

In a report presented at a meeting of the Finno-Ugric Society in Helsinki in 1997, immediately after the publication of the Mator etymological dictionary [9], E.A. Helimski noted that the Mator language should be grouped together with Nenets and Enets languages [10]. This conclusion arose from an analysis of all available wordlists of the Mator language, which demonstrated that the largest number of lexical isoglosses linked the Mator language with Enets and Nenets. Furthermore, they also shared common phonetic and morphological isoglosses. Given the results he had previously obtained by analyzing 92-word lists, E.A. Helimski leaned towards the idea that the classification of Samoyedic languages changed over time due to migrations and areal contacts. As a proposed solution, he introduced a scheme consisting of three periods, as illustrated in Fig. 5.

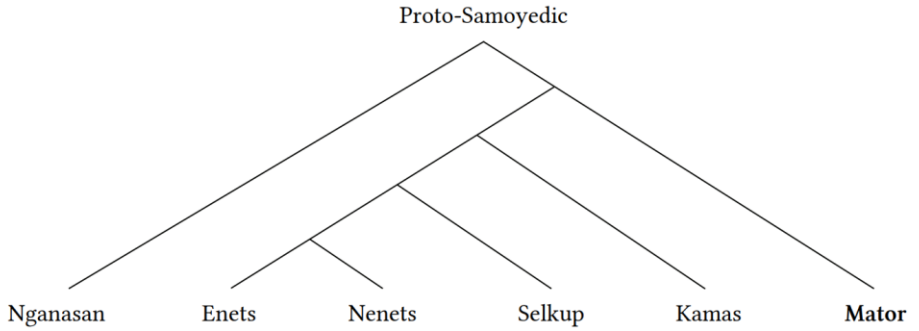

Fig. 4. Classification of Samoyedic languages according to [8].

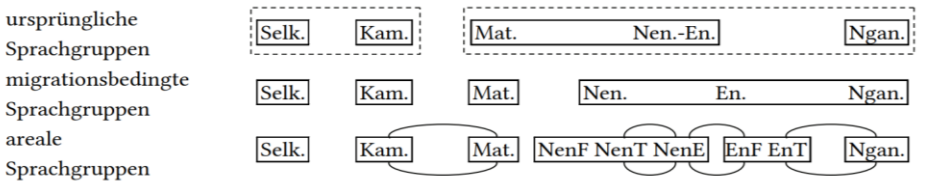

Fig. 5. Classification of Samoyedic languages according to [10].

In Helimski's work, the main argument for the original proximity of Mator to the Nenets-Enets group was the number of shared Proto-Samoyedic etymologies for which there are no reflexes in Kamasin and Selkup. He found 96 such lexemes in Y. Janhunen's dictionary [11], of which 54 have no Nganasan parallels. Meanwhile, according to [11], there are only 34 shared etymologies between Kamasin and Northern Samoyedic languages, excluding Mator and Selkup. In a report in 1997, E.A. Helimski suggested that this could indicate the existence of a Mator-Nenets-Enets group immediately after the divergence of the Proto-Samoyedic community, with 54 shared words being its legacy [10].

However, it should be noted that this work was prepared by E.A. Helimski in 1997 and was not published during the subsequent 10 years of his life. It can be assumed that this was not accidental. The reason could be that the preparation and publication of new sources on Samoyedic languages, including those by Helimski himself on Nganasan [12], Enets [13], Northern Selkup [14-15], and a

large dictionary on Southern and Central Selkup dialects created by the followers of A.P. Dulzon (see [16]) had significantly expanded the possibilities for finding new Samoyedic etymologies. It turns out that, on the one hand, for some Proto-Samoyedic words that, according to [11], had commonly known reflexes only in Nenets, Enets, and Mator languages and were part of the list consisting of 54 lexical isoglosses, reflexes can also be found in Southern and Central Selkup dialects according to [16]. Here are some examples: Proto-Samoyedic **k̂əj-* 'to go' (> Nenets, Mator) [11] can be compared with Selkup Ob', Tym., Vas. *қвая-гу*, Vas. Tym. *қвое-гу, қвоя-гу* 'to go, to ride' [16], Ob', Narym. *қва-гу, қваг-гу* 'to go, to ride' [16]; Proto-Samoyedic **l̂əpta* 'flat, level, plain' (> Enets, Nenets, Mator) [11] can be compared with Selkup. Narym., Tym. *лапчяль* 'small, flat' [16], also see further Selkup dialectal forms in [17].

On the other hand, E.A. Helimski prepared but did not publish a dictionary of Northern Samoyedic nouns and adjectives, which was compiled and supplemented with modern forms by J.V. Normanskaja and V.Y. Gusev. It is currently available online on the LingvoDoc platform at the link [18]. It contains 836 Northern Samoyedic stems, some of which are well-known and have parallels in other Samoyedic languages, including Mator, Kamasin, and Selkup. According to [11], 325 words in this list have such parallels. At the same time, 511 Northern Samoyedic stems do not have commonly known parallels in Selkup, Kamasin and Mator languages. Out of these, 394 lexemes have parallels in Nganasan and Nenets and/or Enets, while 117 have parallels only in Nenets and Enets. It should be noted that these numbers only concern stems of nouns and adjectives, the analysis of verbs would likely increase this count significantly. Thus, it is evident that when analyzing complete dictionaries, the number of shared lexemes that are found exclusively in Nenets, Enets, and Nganasan is significantly higher than in Nenets, Enets, and Mator (394 based only on nouns and verbs and 96 based on the entire Mator dictionary). Certainly, this may be related to the fact that Nganasan dictionary, prepared by E.A. Helimski (see [12]), is much larger than the dictionary of the already-extinct Mator language (see [9]). However, it is clear that the calculations presented in [10] cannot be considered to be a definitive argument for deciding which classification is correct.

In 2016, V. Blažek, in an article [19] dedicated to the memory of E.A. Helimski, turned to the topic of analyzing the basic vocabulary of Samoyedic languages. As mentioned earlier, between 1982 and 2016, a significant number of new, more comprehensive dictionaries of specific Samoyedic languages and etymological dictionaries of those languages were published. Based on these, V. Blažek created new lists of basic vocabulary. The results of the calculations he obtained can be seen in Fig. 6 and Fig. 7.

In this scheme, the maximum and minimum possible dating are indicated on the right and left respectively, depending on which synonym is considered to be the base word.

Below in Fig. 7, the scheme of Samoyedic language divergence is presented as a diagram that reflects the percentages of similarity between all languages.

We can see that the classification proposed by V. Blažek based on glottochronological calculations essentially coincides with the traditional classification (see Fig. 1).

In 2023, A.Y. Urmanchieva in her doctoral dissertation [20], raises the question of why the conclusions presented by E.A. Helimski in his 1997 report [10] about the closeness of the Mator language to Nenets and Enets were not confirmed by the works on glottochronology published by S.A. Starostin [6] and V. Blažek [19]. She also creates lists of basic vocabulary, but for nine Samoyedic languages, adding data on Forest Nenets, Forest Enets, and Southern Selkup languages to the calculations. As a result of the analysis of two lists: M. Swadesh's list and a wordlist obtained in the "Loanwords Typology" project – the so-called Leipzig-Jakarta list, published in [21], A. Y. Urmanchieva concludes that the Mator language formed a special subgroup with the Nenets language, a Proto-Nenets-Mator subgroup. According to the author, in order to classify the languages correctly, four levels of communication between the speakers of those languages must be taken into account: the ancient contacts, the later contacts, the most recent contacts and the historical epoch of contacts. In summary, it can be stated that the conclusions of A.Y. Urmanchieva regarding the classification of

the Samoyedic languages at different time periods are similar to the report by E.A. Helimski in 1997, with the exception that she postulates a closer affinity between the Nenets language and Mator, and not with Enets. Furthermore, while the classification of the Samoyedic languages immediately following the divergence of the proto-language by A.Y. Urmanchieva is based on the analysis of lists of basic vocabulary, it significantly differs from both the initial classification by E.A. Helimski proposed in 1982 [5], also based on lists of basic vocabulary, and the results by S.A. Starostin (see Fig. 2) and V. Blažek (see Fig. 6, 7), obtained from materials of contemporary dictionaries. The numbers in the Fig. 6 indicate the date of the Proto language split. Negative numbers indicate that the time of the splitting was BC. The percentages show how many percent of the words from the hundred-word list have common origin. The question arises as to how several authors could obtain such different results when applying the same method?

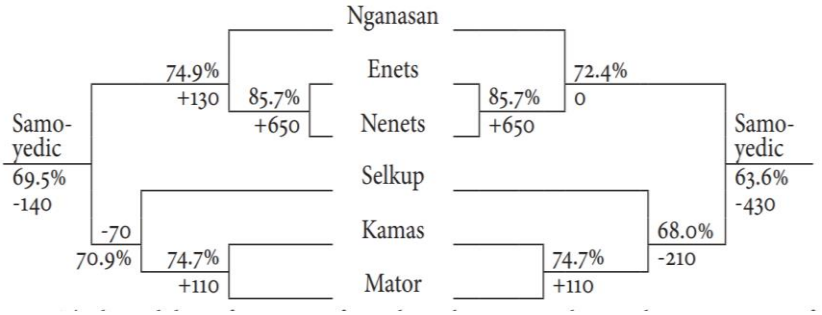

Fig. 6. Time of divergence of Samoyedic languages according to [16].

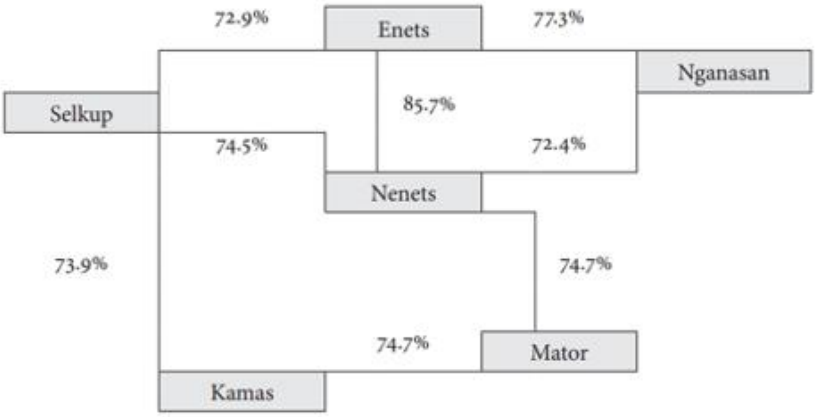

Fig. 7. Percentage of shared basic vocabulary in Samoyedic languages according to [16].

According to E.A. Helimski's calculations, the percentage of correspondences between Tundra Nenets and Mator is 58.1%, while between Tundra Nenets and Tundra Enets, it is 74.7%. According to V. Blažek, the correspondence between Tundra Nenets and Mator is 74.7%, and between Tundra Nenets and Tundra Enets, it is 85.7%. However, according to A.Y. Urmanchieva, the correspondence between Tundra Nenets and Mator is 74.19%, and between Tundra Nenets and Tundra Enets, it is 69.23%. Is this related to the fact that A.Y. Urmanchieva used a different approach in analyzing the 100-word dictionary of Mator? In particular, "when synonyms expressing the same meaning are present, the presence of a cognate in another Samoyedic language for a Mator word is noted, but the absence of a cognate for one of the synonyms in Mator is not noted" (see more in [20]). She indicates that this method yields "higher percentages of correspondence in the Mator column, so these data can only be compared with other data within this column, but not with other columns". However, the question arises as to how to compare the similarity figures between Tundra Nenets and Tundra

Enets and between Tundra Nenets and Mator in this case. Can one postulate the existence of a special Nenets-Mator group based on results obtained by different methods? Besides the slightly modified method used by A.Y. Urmanchieva, uses different lists of basic vocabulary, all of which are available online: those of E.A. Helimski, V. Blažek, A.Y. Urmanchieva, and Y.B. Koryakov, who in 2018 analyzed the similarity of only Nenets, Enets, and Selkup languages (see [22]). Firstly, A.Y. Urmanchieva removed words from M. Swadesh's list, such as 'bird' (lacking a single-word translation, it is described as 'flying animal'), 'seed' (weakly represented), 'leaf' (weakly represented), 'feather' (difficulties in choosing a neutral term), 'green' (lack of a stable translation for this color), 'yellow' (lack of a stable translation for this color), 'round' (difficulties in choosing a neutral term), see [22]. Secondly, for the Mator language, for example, our knowledge of its lexicon was significantly supplemented by the dictionary [9], which includes some words that were absent from the lists of basic vocabulary published in [5], and the etymologies of of all words were clarified. For instance, E.A. Helimski provides the Mator word *тчелегадe* meaning 'root', and his 1982 work does not indicate that this is a Khakas loanword, see [9]. Additionally, the Mator dictionary [9] provides another word for 'root' that has a Proto-Samoyedic etymology PS **wanco* 'root' [11], the reflexes of which are Tundra Nenets *вано* and Tundra Enets *badu* ('root'). Consequently, such cases allowed A.Y. Urmanchieva, V. Blažek, and S.A. Starostin, to increase the percentage of similarity between Nenets and Mator languages compared to the study by [5]. However, upon careful examination of the lists, it becomes clear that the differences in them are not solely due to the use of more recent dictionaries. We will illustrate this with the first word from the list of basic vocabulary and only three languages: Nenets, Enets, and Selkup, which are present in four works:

‘all’:

[5]: Nenets *тюку’*, Enets *ťuku’,* Selkup. *muntį̈k.*

[19]: Nenets *ťuku’, ťübea, ŋokxa, mal’,* Enets *ťuku’, kurhari*. Baj Selkup, Taz *mûndeŋ*, Upper Ob, Chaya *sielaŋ*, Middle Ob, Ket, Pumpokol *wes*.

[22]: Forest Nenets *чупей* ‘all, complete; completely’, Tundra Nenets: *сямян* ‘all’, *мал’* ‘all’; Two or three different roots. Tundra Enents *t’uku*, *t’ukut’ii*; *tʃuku*, *tʃukotʃii*. Forest Enets *tʃuku*, *tʃuktʃii.* Selkup Taz *muntį̈ŋ* ‘all, everybody’; *мунтык/ӈ* ‘all, everybody’, Narym, Ob Chumylkup *вес / вэс,* Ket *tīr* ‘all, everybody’.

[20]: Tundra Nenets *tyuku°q*, Tundra Enets. *t’ukut’ii*, *t’uku*, Selkup North. *мунтык/ӈ*, South. *tāķ əl’*.

It becomes evident that among the considered lists, the data used by V. Blažek differ the most. This is because he combines data from all available dictionaries and dialects in a single column. Furthermore, he does not conduct calculations separately for each dialect but treats forms from different dialects as synonyms. Therefore, it's apparent that the percentage of common words for different languages is significantly higher than that of other researchers.

We can observe in this example that the lists of E.A. Helimski and A.Y. Urmanchieva are similar to each other. This tendency generally holds for other words as well, except for the fact that A.Y. Urmanchieva also uses materials from three additional languages (Forest Nenets, Forest Enets, South Selkup) and takes into account materials from more recent dictionaries.

The difference between the lists of E.A. Helimski and A.Y. Urmanchieva compared to the materials of Y.B. Koryakov in this example lies, firstly, in the choice of Tundra Nenets word. In this case, it is impossible to determine whose list is more accurate without conducting surveys of native speakers since all three words in the dictionary [22], *мал*’, *сямян*, and *тюку*’’ seem to have the meaning 'all' based on their contexts. Regarding South Selkup, different forms are also listed in Y.B. Koryakov's and A.Y. Urmanchieva's lists, e.g., [22]: *tīr,* [20]: South Selkup. *tāķ əl’*. However, in the dictionary [16], which is based on a card index created by A.P. Dulzon and his students, and includes over 25,000 lexemes with contexts, the most common South Selkup (Ket) word for 'everything, all' is a Russian borrowing *вес*, and the same is indicated only in V. Blažek’s list.

Undoubtedly, in a situation when extensive dictionaries containing numerous synonyms but lacking easily accessible corpora or a literary norm, it is necessary to conduct a survey of native speakers to make a more reliable choice for the word in the Swadesh list. However, only in the work of Y.B. Koryakov [22], it is mentioned that the material was not collected from existing dictionaries but rather through surveys of speakers, which were conducted for the Nenets language by M.K. Amelina, for the Enets language by A.B. Shluinsky and O.V. Khanina, and for the Narym Selkup language by N.L. Fedotov [1]. Certainly, E.A. Helimski and A.Y. Urmanchieva also relied on their own field materials, but it is not always clear whether a lexeme in their lists was taken from dictionaries or obtained through surveys of speakers.

Therefore, in further work, we will be relying on word lists presented online on the LingvoDoc platform [1] collected from native speakers and already connected by etymological links.

3. Analysis of Samoyedic languages' basic vocabulary lists using the LingvoDoc platform

Currently, on LingvoDoc, any user can create their own dictionaries and/or corpora and analyse those created by other users, if they were made available to the public. In dictionaries, under the "Tools" tab, there is an option for "Glottochronological analysis of languages". This option can be applied to any set of languages in the dictionaries of which more than 50 words from M. Swadesh's Swadesh list are present. The 100-word Swadesh - Starostin list [23] is chosen because it is the only one for which strict semantic specifications have been developed and justified [23-24], allowing for sufficiently precise analysis of comparable data for different languages.

For many words in the Swadesh list, speakers of Samoyedic languages provide several synonyms. For a more precise procedure, a survey was conducted using a questionnaire with sentence examples for each word [25].

According to the proposed glottochronology by S.A. Starostin, see more details [26], starting with the calculation using the formula presented above, loanwords are first removed, then related words are connected on the LingvoDoc platform by etymological links, after that the percentage of matches between the lists of two idioms is calculated, and the time of divergence is determined [26].

This formula, for which S.A. Starostin found an experimentally determined lambda of 0.05, allows for determining the time of divergence of any set of languages, see more [26].

LingvoDoc also includes a feature for creating language similarity graphs in 2D and 3D formats.

Currently, on LingvoDoc, basic vocabulary lists are available for:

1) Tundra Nenets, created based on the dictionary [27] and a survey of speakers from the Yamal Tundra conducted by M.K. Amelina in 2017.
2) Nenets, created based on the dictionary by A.A. Dunin-Gorkavich, 1910.
3) Forest Nenets, created based on a survey conducted by M.K. Amelina in the village of Khalyasavey in 2015.
4) Tundra Enets, created in 2015 based on a survey in Dudinka by O.V. Khanina.
5) Forest Enets, created in 2011 based on a survey in Dudinka by O.V. Khanina.
6) Nganasan, created in 2022 based on a survey in Dudinka.
7) Taz Selkup, created in 2022 based on a survey in Krasnosel'kup.
8) Surgut (Taz) Selkup, collected by P.S. Pallas in the 18th century.
9) Taz Selkup, created based on the dictionary [14] and a survey of speakers conducted by O.A. Kazakevich in the village of Bystrika in 2011.
10) Ket Selkup, collected by P.S. Pallas in the 18th century.
11) Ket Selkup, recorded by E.M. Budyanskaya and O.A. Kazakevich in the village of Katayga in 2010.

12) Lower Chaya Selkup, created based on the works of N.P. Grigorovsky from 1879.
13) Narym Selkup, created based on recordings by Y.A. Morev, made in the village of Laskino in 1967-1970.
14) Narym Selkup, created based on recordings by N.L. Fedotova, made in the village of Parabel in 2009.
15) Mator language, created based on the dictionary [9].
16) Kamas Selkup, created based on the dictionary [28].
17) Kamas concordance based from Kamas corpus glossed in INEL project [29].

The glottochronological analysis of these lists resulted in the following 3D graph representing the degree of proximity in the divergence time of Samoyedic languages and dialects, see Fig. 8.

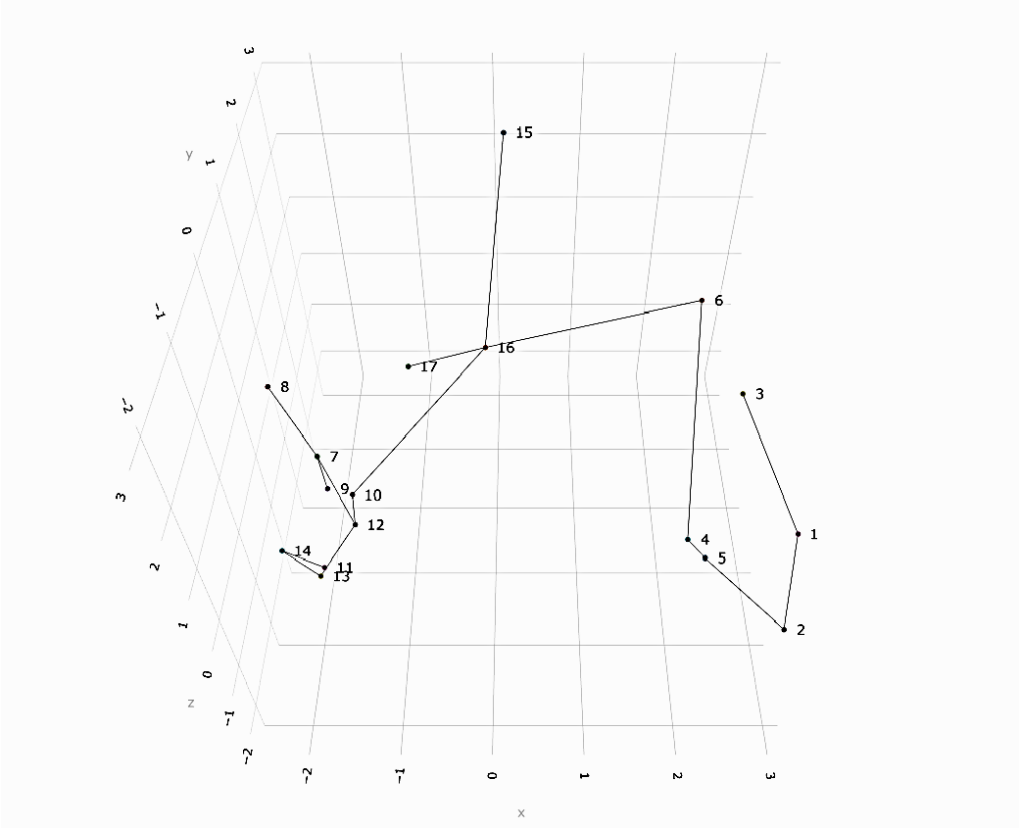

Fig. 8. Graph of the degree of proximity in the divergence time of Samoyedic languages.

In digital format, the obtained results are as follows, see Table 1. The first digit in the table represents the time in millennia that has passed since the divergence of two languages or dialects. The second digit reflects the percentage of etymologically related words that match. From the table, it can be observed that when analyzing complete Swadesh lists (excluding monuments from the 18th century, where a lower percentage of similarities is often observed, possibly due to the the incomplete state of those lists), the lowest percentage of similarities (55-56%) in words is found between Nganasan and Southern and Central Selkup dialects. According to S.A. Starostin's formula, they diverged approximately 2700-2800 years ago, i.e., in the early 1st millennium BCE. This percentage is higher than that of A.Y. Urmanchieva (between Nganasan and Selkup – 40.8%), see [20], and significantly lower than those of V. Blažek (63.6-69.5%, numbers represent the results of different synonym calculations). This is likely related to the principles of selecting basic vocabulary words. The obtained

result is nearly identical to the percentage obtained by E.A. Helimski in 1982 (between Nganasan and Selkup – 54.6%) and with the dating of the Proto-Samoyedic language divergence proposed by S.A. Starostin (see Fig. 2).

Based on the obtained results, we will attempt to answer the questions raised by other researchers:

1) **Did Northern Samoyedic language group, consisting of Nenets, Enets, and Nganasan languages exist?** Or, as claimed by E.A. Helimski in the 1997 report, is the Mator language closer to Nenets and Enets than Nganasan? It turns out that while the proximity of Mator to Nenets is slightly higher (65-67%) than that of Nganasan (63-66%), there is a significant difference in proximity with Enets: Mator with Enets (66-68%) and Nganasan with Enets (72-75%). Therefore, according to S.A. Starostin's methodology, it is reasonable to postulate a Northern Samoyedic language group. Is it possible that the proximity between Nganasan and Enets languages has a contact nature? Let's examine which words in the basic vocabulary form Nganasan-Nenets-Enets isoglosses to understand their nature, whether they are reflexes of Proto-Samoyedic roots or the result of later contacts.

   Nganasan-Nenets-Enets parallels in the Swadesh list with no corresponding lexemes in the basic vocabulary of other languages, see Table 2.

Table 1. The time of diversity and the percent of the common words in Samoyed languages.

	1. Tundra Nenets	2. Nenets 1910	3. Forest nenets	4. Tundra Enets	5. Forest Enets	6. Ngana san	7. Taz Selkup	8. Taz Selkup XVIII century	9. Upper Tolka Selkup	10. Ket Selkup XVIII century	11. Ket Selkup	12. Lower Chaya 1879 г.	13. Laskino Selkup 1967	14.Narym Selkup	15. Kamasin Selkup	16. Mator language
1	n/a	1.17 (87%)	1.32 (85%)	1.89 (73%)	1.61 (79%)	2.40 (63%)	2.53 (60%)	2.61 (59%)	2.71 (57%)	2.39 (63%)	2.54 (60%)	2.39 (63%)	2.61 (59%)	2.59 (59%)	2.54 (60%)	2.29 (65%)
2	1.17 (87%)	n/a	1.79 (75%)	1.92 (73%)	1.67 (78%)	2.50 (61%)	2.72 (57%)	2.67 (58%)	2.71 (57%)	2.78 (56%)	2.83 (55%)	2.29 (65%)	2.84 (54%)	2.75 (56%)	2.55 (60%)	2.67 (58%)
3	1.32 (85%)	1.79 (75%)	n/a	2.10 (69%)	1.82 (75%)	2.23 (66%)	2.53 (60%)	2.51 (61%)	2.48 (61%)	2.45 (62%)	2.48 (61%)	2.39 (63%)	2.44 (62%)	2.53 (60%)	2.21 (67%)	2.21 (67%)
4	1.89 (73%)	1.92 (73%)	2.10 (69%)	n/a	0.92 (92%)	1.82 (75%)	2.34 (64%)	2.26 (66%)	2.51 (61%)	1.85 (74%)	2.50 (61%)	2.23 (66%)	2.49 (61%)	2.49 (61%)	2.19 (67%)	2.15 (68%)
5	1.61 (79%)	1.67 (78%)	1.82 (75%)	0.92 (92%)	n/a	1.93 (72%)	2.46 (62%)	2.51 (61%)	2.60 (59%)	2.20 (67%)	2.61 (59%)	2.30 (65%)	2.51 (61%)	2.59 (59%)	2.25 (66%)	2.25 (66%)
6	2.40 (63%)	2.50 (61%)	2.23 (66%)	1.82 (75%)	1.93 (72%)	n/a	2.44 (62%)	2.39 (63%)	2.66 (58%)	2.16 (68%)	2.80 (55%)	2.60 (59%)	2.84 (55%)	2.78 (56%)	2.21 (67%)	2.17 (67%)
7	2.53 (60%)	2.72 (57%)	2.53 (60%)	2.34 (64%)	2.46 (62%)	2.44 (62%)	n/a	1.23 (86%)	0.89 (92%)	1.04 (90%)	1.13 (88%)	1.14 (88%)	1.29 (85%)	1.31 (85%)	2.19 (67%)	2.21 (67%)
8	2.61 (59%)	2.67 (58%)	2.51 (61%)	2.26 (66%)	2.51 (61%)	2.39 (63%)	1.23 (86%)	n/a	1.47 (82%)	1.71 (77%)	1.75 (76%)	1.64 (78%)	1.66 (78%)	1.85 (74%)	1.77 (76%)	1.94 (72%)
9	2.71 (57%)	2.71 (57%)	2.48 (61%)	2.51 (61%)	2.60 (59%)	2.66 (58%)	0.89 (92%)	1.47 (82%)	n/a	1.74 (76%)	1.41 (83%)	1.10 (89%)	1.33 (85%)	1.43 (83%)	2.13 (68%)	2.36 (64%)
10	2.39 (63%)	2.78 (56%)	2.45 (62%)	1.85 (74%)	2.20 (67%)	2.16 (68%)	1.04 (90%)	1.71 (77%)	1.74 (76%)	n/a	1.18 (87%)	0.99 (91%)	1.37 (84%)	1.18 (87%)	1.82 (75%)	1.80 (75%)
11	2.54 (60%)	2.83 (55%)	2.48 (61%)	2.50 (61%)	2.61 (59%)	2.80 (55%)	1.13 (88%)	1.75 (76%)	1.41 (83%)	1.18 (87%)	n/a	1.11 (89%)	0.98 (91%)	0.91 (92%)	2.38 (63%)	2.00 (71%)
12	2.39 (63%)	2.29 (65%)	2.39 (63%)	2.23 (66%)	2.30 (65%)	2.60 (59%)	1.14 (88%)	1.64 (78%)	1.10 (89%)	0.99 (91%)	1.11 (89%)	n/a	0.95 (91%)	1.06 (89%)	2.01 (71%)	1.95 (72%)
13	2.61 (59%)	2.84 (54%)	2.44 (62%)	2.49 (61%)	2.51 (61%)	2.84 (55%)	1.29 (85%)	1.66 (78%)	1.33 (85%)	1.37 (84%)	0.98 (91%)	0.95 (91%)	n/a	0.97 (91%)	2.33 (64%)	2.10 (69%)
14	2.59 (59%)	2.75 (56%)	2.53 (60%)	2.49 (61%)	2.59 (59%)	2.78 (56%)	1.31 (85%)	1.85 (74%)	1.43 (83%)	1.18 (87%)	0.91 (92%)	1.06 (89%)	0.97 (91%)	n/a	2.33 (64%)	2.17 (67%)
15	2.54 (60%)	2.55 (60%)	2.21 (67%)	2.19 (67%)	2.25 (66%)	2.21 (67%)	2.19 (67%)	1.77 (76%)	2.13 (68%)	1.82 (75%)	2.38 (63%)	2.01 (71%)	2.33 (64%)	2.33 (64%)	n/a	1.74 (76%)
16	2.29 (65%)	2.67 (58%)	2.21 (67%)	2.15 (68%)	2.25 (66%)	2.17 (67%)	2.21 (67%)	1.94 (72%)	2.36 (64%)	1.80 (75%)	2.00 (71%)	1.95 (72%)	2.10 (69%)	2.17 (67%)	1.74 (76%)	n/a

Table 2. Nganasan-Nenets-Enets parallels in the Swadesh list with no corresponding lexemes in other Samoyed languages.

Nganasan	Tundra Nenets	Nenets 1910	Forest Nenets	Tundra Enets	Forest Enets	Proto Samoyed
1. Head нгойбу́	Head ӈэва	Head айво	Head ŋǽ̆iwa	Head aburʲe, aburʲi	Head ɛba	*åjwå Head < Proto Uralic*ojwa
2. To burn лонгу́ди				Burn leided'e		
3. Yellow тодяку́	Yellow тасехэй	Yellow тасей				
4. Stomach ми́на̇да̇			Stomach mɨntɕɨ́	Stomach munzɨ, munðe, munðɨ	Stomach Munoδe	
5. Star фотади́е				Star poδeseo	Star poδesei	
6. Dry kośüa	Dry хасуй	Dry хазуй	Dry kɨsúj	Dry kasuo	Dry Kasuj	*kås(ə̂)- to become dry < Proto Uralic *kośkɨ-
7. Human нгана́са	Human ненэць’(н)	Human нена́чь	Human ɲéɕaŋ		Human entʃiʔ	*enå real
8. What ма'				What mʲi, mʲiʔ, mʲiiʔ		*me what < Proto Uralic *mз
9. Lay down bəd-	Lay down ва”нă(сь)			Lay down baδoti	Lay down balʲkiʃ	

So, out of 9 lexical isoglosses, 5 words (to burn, yellow, stomach, star and lay down) do not have an etymology beyond the Northern Samoyedic languages. However, the remaining words that account for 72-75% similarity between Enets and Nganasan either have Proto-Samoyedic etymology or parallels in other Southern Samoyedic languages. Therefore, we assume that the Northern Samoyedic group cannot be entirely the result of later contacts. It is evident that this group existed briefly after the divergence of the Proto-Samoyedic language around the mid-1st millennium BCE. There were likely separate contacts between Enets and Nganasan, possibly involving a substrate language.

2) **Did a Southern Samoyedic commonality exist, uniting Mator, Kamasin, and Selkup languages?** Or is the Mator-Kamasin group closer to Northern Samoyedic than it is to Selkup, as proposed by S.A. Starostin?

The percentage of word matches with modern Selkup dialects for Kamasin language is 63-68%, and for Mator language, it's 64-71%. However, when considering the basic vocabulary from the 18th-19th centuries, the percentage of their overlap with Selkup dialects is much higher: Kamasin (64-76%) and Mator (69-75%).

The overlap between Kamasin and Northern Samoyedic languages is comparable to modern Selkup dialects: with Nenets 60-67%, with Enets and Nganasan 67%. Mator also exhibits a similar percentage of overlap with Northern Samoyedic languages: with Nenets 65-67%, with Enets and Nganasan 66-68%.

Let's consider specific words that constitute Kamasin-Mator-Selkup isoglosses. These are the reflexes of Proto Samoyed **olə̂* head, **perkä* belly, **kuńćз* star, **tipå* man, **mə̂jå* earth, **si*$_1$*nsä*$_1$*kkз* bird [11]. Just one word has just Selkup-Kamass parallel: Taz Selkup *ippiqo* lie, Kamass *i'büm* lie. So, it is evident that 6 out of 7 words are reflexes of Proto-Samoyedic lexemes. Thus, the hypothesis of a later contact influence can be ruled out, and the existence of a Southern Samoyedic group can be confidently acknowledged, supported by materials from the early Selkup dictionaries and books.

3) **Between which Samoyedic languages did long-lasting commonalities exist?** Our analysis allowed us to identify only those commonalities that most researchers have recognised. A clear increase in the percentage of shared words is observed between:
   a) Nenets and Enets languages (69-79% similarity, dating their divergence to the 1st-5th centuries AD).
   b) Mator and Kamasin (76% shared words), with the divergence dating back to the 3rd century AD.
   c) Southern and Northern Selkup languages (74-92%); due to the wide range of the similarity percentage, it is challenging to reliably date the time of divergence.

The close relationship between Tundra and Forest Nenets (85% shared words) is indisputable, with their divergence dating to the 5th century AC. These languages were once considered a single language by some authors. Later differences, such as those between Tundra and Forest Enets, have a dialectal character, as these idioms share 92% of their words and, according to S.A. Starostin, diverged in the 2nd millennium AC.

Let's examine which specific lexemes highlight the commonalities mentioned above: Nenets-Enets (see Table 3), Mator-Kamasin (see Table 4), and Selkup (see Table 5).

Table 3. Nenets-Enets parallels in the Swadesh list with no corresponding lexemes in the basic vocabulary of other languages.

Tundra Nenets	Nenets 1910	Forest Nenets	Tundra Enets	Forest Enets	Proto Samoyed
1. Road, path сехэры	Road, path се́ры		Road, path *sexa*	Road, path sɛxeri	Proto Uralic *ćäke(-rɜ) Hard snow
2. Smoke якэ	Smoke яко́	Smoke dʲakí		Smoke dʲake dʲakʲi	*jäckə̂ - Smoke
3. Green пăдярăха			Green poðeraxa	Green pɔzeraxa	*pə̂tä bile < ПУ *piša bile
4. Sand тăб	Sand табъ		Sand tobo	Sand tɔb(u)	*t1ə̂pə̂ Dirt, soil
5. Black пăриденя	Black парде́на	Black plʲit͡ɕéna		Black pɔlzer-	
6. This тюку		This t͡ɕukʲí	This tʃĭkʲi, tʃĭkʲe, tʃĭke	That tʃĭkʲi, tʃĭkʲe, tʃĭk	*tü- That
7. Tongue нямю	Tongue ня́мю		Tongue nʲamʲi	Tongue nʲamʲi	*äŋ Mouth + *mü inside
8.	Stomach се́нзе			Stomach sɔse	*sünsə̂ chest < Proto Uralic *śińćä
9.	Road, path сарпе́			Road, path sɔbʲi	*sə̂rpä path
10.	Small о́лику		Small ulʲajku, ulʲajguu, ulʲajgu	Smallibl̩ʲe:juk, ibl̩ʲe:jku, ibl̩ʲa:jku	
11.		Round pǽɛku	Round pos'uteo	Round pɔʃtej	

Table 4. Mator-Kamasin parallels in the Swadesh list with no corresponding lexemes in the basic vocabulary of other languages.

Mator	Kamasin	Proto Samoyed
1. Road, path ād'e	Road, path иде	*etå path, footsteps
2. Sand phüra	Sand хура	

Table 5. Selkup words in the Swadesh list with no corresponding lexemes in the basic vocabulary of other languages.

Taz Selkup.	Taz Selkup XVIIIв	Upper Tolka Selkup	Narym district (Ket)XVIIIв .	Ket Selkup.	Lower Chaya Selkup. 1879	Laskino Selkup 1967 and 1970 гг.	Narym selkup.	Proto Samoyed
1. To burn *č́ɔpi̯qo*		To burn *ʧʲopa-*						
2. Road, path wḙtti̯		Road, path *vətti*		Road, path vəttɨ	Road, вато́нъ	Road, wát	Road, path vat	*uə̂t footstep, path < Proto Uralic *utka
3. Small *ki̯pa*	Small *кыбыля*	Small *kɨpa*	Small кыба		Small кыба́	Small qɨba	Small kɨbá	
4.Lot *koč́č́i̯*		Lot *koʧʧʲ*		Lot kwəʧʧi	Lot ко́ чи	Lot qoč'ə, kóč'ə, koč'	Lot kɔ́tɕ(ik)	
5. Man *ira*		Man *ɨra*						
6. New *š́ɛnti̯*		New ʃʲentɨ		New sʲendɨ	New се́нды	New ʂénd, ʂänd̂	New ʃand	
7. Leg (foot) topi̯	Leg (foot) то- пы	Leg (foot) topɨ	Leg (foot) то- ба	Leg (foot) topɨ	Leg топъ (foot)	Leg (foot) tob	Leg (foot) tɔb	*topå claw, hoof
8. nose i̯ ntäl	Nose унжель	Nose inʲʧʲæj						
9.Sand kōra		Sand qora		Sand kora	Sand кора́	Sand qóːra	Sand kɔ́ra	
10. Dry tḙki̯pi̯			Dry tøkɨmpan		Dry ʧəkɨbʲi		Dry tɕágəbəl	
11. Ear üŋki̯ lsa	Ear юнголсотъ							*jüntə̂ - hear
12.human qum	Human кубъ	Human qum	Human куммъ	Human qum	Human гомъ, комъ	Human qop, qup	Human kup	Proto Uralic *koj(e)mз
13. what qaj	What кае		What кай	What qaje	What ка́ й	What qái-l, qáj-t		Tungus *xai
14. Chest ki̯ li̯		Chest kɨlʲɨ		Chest kɨlɨ		Chest qɨ́l	Chest kɨl	
15. Full tiri̯ k				Full tɪrɨmbʲitɨj			Full tɨr	*tirə̂ ~ *terə̂ contents
16. To say kḙ ti̯ qo		To say kət-		To say qət-	To say ка́ с-, ка́ т-	To say qadə-gú	To say kadəgú	*kḙt(ə̂)- To say
17.	Egg иѓа	Egg ɛŋʲ						Middle Ket ɛʔj egg, ēŋ eggs
18.				Egg nʲapʲi		Egg n'áb'i	Egg n'ab'	
19.	To eat еганъ		To eat еганъ					
20.			White чаягасомба	White tʲɛkɨ	White те́гы	White č'áĝ čágə, č'aːR,		White tɕ'ag
21.				Woman pəjja			Woman pajá	
22.					This та́б, та́у	This táu, táw	This tav	*tə̂- This

We have identified only 2 isoglosses highlighting the Kamasin-Mator community. Moreover, the high percentage of similar lexemes (76%) has parallels in various languages, indicating that these isoglosses are Proto-Samoyedic and cannot be explained by later contacts.

Furthermore, in each Selkup dialect, there are 3 to 18 words with unknown etymology. It is evident that a significant number of Selkup words that lack parallels in other Samoyedic languages serve as evidence for the close relationship of this group and a lengthy period of its separate development. The abundance of words without reliable etymology suggests the possibility of substrate influence.

4. Conclusion

The conducted analysis has allowed us to identify four types of increases in the percentage of similarity when analysing the basic vocabulary among Samoyedic languages.

Type 1 is observed between Mator and Kamasin languages and between Selkup, Mator, and Kamasin languages. The number of unique isoglosses between these languages, absent in other Samoyedic languages and dialects, is small: 2-7 lexemes, most of which belong to the Proto-Samoyedic basic vocabulary. The increased percentage primarily results from the preservation of words from the Proto-Samoyedic basic lexicon, which had been lost in some other Samoyedic languages.

Type 2 of lexical isoglosses is observed among North Samoyedic languages according to traditional classification. There are 9 words of basic vocabulary shared between Nganasan, Enets, and Nenets, of which 5 have no etymology. This correlates with the observation that the Nganasan basic vocabulary contains a considerable number of words with no etymology, as noted by all researchers see more [20]. According to our calculations, there are 13 of them. Meanwhile, in Nenets, Enets, Kamasin, and Mator languages, the divergence of which dates back to our era, there are no more than 5-6 lexemes.

Type 3 of lexical isoglosses is observed between Nenets and Enets languages. They share 11 lexical isoglosses in their basic vocabulary, with 3 words having no etymology. Most of the words in this category have Proto-Samoyedic or even Proto-Uralic etymology but reconstructed meanings that do not belong to the basic vocabulary. Thus, in 8 cases, shared semantic innovations occurred in Nenets and Enets languages, confirming the existence of a Nenets-Enets group that separated in the early 1st millennium AD.

Type 4 of lexical isoglosses is observed among Selkup languages. Out of 22 words, only 2 ('this,' 'to say') belong to the Proto-Samoyedic basic vocabulary, 5 words have changed meanings, and 15 words are either presumed ancient borrowings or have unknown etymology. Considering that Narym Selkup has 13 words with unknown etymology, with no parallels even in Selkup dialects, it is possible to assume both prolonged separate development and significant external influence on Selkup.

In conclusion, the analysis supports the traditional classification of Samoyedic languages. The integration of material from the first Selkup books into the LingvoDoc platform allowed for a more reliable postulation of the South Samoyedic group.

The integration of S.A. Starostin's formula on to the LingvoDoc platform has allowed us to approximate the times of divergence for various Samoyedic language groups:

1) Proto-Samoyedic language: Around the beginning of the 1st millennium BCE.
2) North Samoyedic languages: Around the middle of the 1st millennium BCE.
3) South Samoyedic language: Around the end of the 1st millennium BCE.
4) Nenets-Enets group: Around the beginning of the 1st millennium AD.
5) Kamasin-Mator group: Around the middle of the 1st millennium AD.
6) Selkup group: Around the end of the 1st millennium AD.
7) Nenets language: Around the end of the 1st millennium AD.
8) Enets language: Around the beginning of the 2nd millennium AD.

The analysis of the material indicates that before the beginning of AD, the Samoyedic language groups, both northern and southern, existed, but for a very short period. Nevertheless, the high percentage of similarities between languages within these groups cannot be solely explained by contact influences. Lexical analysis certainly cannot consider a conclusive result in such a sophisticated case, and further confirmation is required through a detailed analysis of morphology and graphic-phonetics.

References

[1]. https://lingvodoc.ispras.ru, дата обращения 10.03.2025.
[2]. Starostin S.A. Comparative-historical linguistics and lexicostatistics // Linguistic reconstruction and ancient history of the East. Ch. I. M., 1989: 3-39.
[3]. Collinder B. Comparative grammar of the Uralic languages. Stockholm, 1960.
[4]. The Oxford Guide to the Uralic Languages, ed. by Bakró-Nagy M., Laakso J., Skribnik E. Oxford, New York 2022.
[5]. Helmski E. The oldest Hungarian-Samoyed language parallels. M., 1982.
[6]. Starostin S. Preliminary results of application of 'recalibrated' glottochronology to classification of Eurasian language families. Paper presented at Workshop on Prehistoric Chronology: Language, Genes and Migrations being held at the Santa Fe Institute, March 2004.
[7]. Katz H. Zur Phonologie des Motorisch-Karagassisch-Taigischen // K. Redei (ed.). Studien zur Phonologie und Morphonologie der uralischen Sprachen. Akten der dritten Tagung für uralische Phonologie, Eisenstadt, 28. Juni —1. Juli 1984 (Studia uralica 4). Wien: Verband der wissenschaftlichen Gesellschaften Osterreichs, 1987. S. 336-348.
[8]. Janhunen J. Samoyedic / The Uralic Languages, ed. by Abondolo, Daniel Routledge. London, 1998, p. 457–479.
[9]. Helimski E. Die matorische Sprache. Szeged: Studia Uralo-Altaica 41, 1997.
[10]. Helimski E. Mator and Northern Samoyed. Materials to a paper read on May 16, 1997 for the Finno-Ugric Society, Helsinki. = Eugen Helimski. Zur Stellung des Matorischen innerhalb der samojedischen Sprachen // V. Gusev, A. Urmanchieva, A. Anikin (eds.). Siberica et Uralica: In memoriam Eugen Helimski (Studia uralo-altaica. Vol. 56). Szeged: University of Szeged, 2022. C. 479–495.
[11]. Janhunen J. Samojedischer Wortschatz. Castrenianumin toimitteita 17. Helsinki, 1977.
[12]. Helimski E. Nganasan dictionary, manuscript. Hamburg, 2007.
[13]. Helimski E. Enets dictionary, manuscript. Hamburg, 2007.
[14]. Helimski E. Selkup dictionary, manuscript. Hamburg, 2007.
[15]. Helimski E., Kahrs U. Nordselkupisches Wörterbuch von F. G. Mal′cev (1903). Hamburg, 2001.
[16]. Selkup-Russian dialect dictionary. Edited by V. V. Bykonya. Tomsk, 2005.
[17]. Alatalo J. Sölkupisches Wörterbuch. (Lexica Societatis Fenno-Ugricae, XXX.) Helsinki: Société Finno-Ougrienne, 2004.
[18]. http://lingvodoc.ru/dictionary/7386/1/perspective/7386/2/view , дата обращения 10.03.2025.
[19]. Blažek V. On the classification of the Samoyedic languages. Dedicated to the memory of Eugene Helimski (1950–2007) // Finnisch-Ugrische Forschungen 63, 2016, s. 79 – 125.
[20]. Urmanchieva A.Y. Reconstruction of the linguistic landscape of Western Siberia (based on the material of Samoyed languages). Dissertation for the degree of Doctor of Philology. M., 2023.
[21]. Tadmor U. Loanwords in the world's languages. Findings and results. In: Haspelmath M., Tadmor U. (eds.) Loanwords in the World's Languages: A Comparative Handbook. Berlin: Mouton de Gruyter, 2009. pp. 55–75.
[22]. Koryakov Yu.B. The problem of "language or dialect" or Samoyedic languages // Ural-Altaic studies 4 (31), 2018, pp. 156 – 217.
[23]. Starostin S. A. Works on linguistics. Moscow, 2007.
[24]. Starostin G. Preliminary lexicostatistics as a basis for language classification: A new approach. Journal of Language Relationship 3, 2010, p. 79-116.
[25]. Kassian A., Starostin G., Dybo A., Chernov V. The Swadesh wordlist. An attempt at semantic specification // Journal of Language Relationship, No. 4 (2010), p. 46-89.

[26]. Starostin G. Lexicostatistics as a basis for language classification: increasing the pros, reducing the cons. In H. Fangerau and H. Geisler and Th. Halling and W. Martin (eds.), Classification and Evolution in Biology, Linguistics and the History of Science: Concepts - Methods - Visualization, Stuttgart, 2013, p. 125- 146.
[27]. Tereshchenko N. M. Nenets-Russian dictionary. M., 1965.
[28]. Donner K. Kamassisches Wörterbuch nebst Sprachproben und hauptzügen der Grammatik, bearbeitet und herausgegeben von A.J. Joki. Helsinki, 1944.
[29]. https://lingvodoc.ispras.ru/dictionary/3596/37747/perspective/3596/37750/view, дата обращения 10.03.2025.

Информация об авторах / Information about authors

Юлия Викторовна НОРМАНСКАЯ – доктор филологических наук, заведующий лабораторией Лингвистические платформы Института системного программирования РАН с 2019 года, ведущий научный сотрудник Институт языкознания РАН с 2003 года. Сфера научных интересов: компьютерные методы в лингвистике, сравнительно-историческое языкознание, уральские языки.

Julia Viktorovna NORMANSKAJA – Dr. Sci. (Philology), Head of the Laboratory Linguistic Platform of the Institute for System Programming RAS since 2019, leading scientist of Institute of Linguistic RAS since 2003. Research interests: computer methods in linguistic, comparative-historical linguistic, Uralic language.