EasyChair Preprint

№ 7446

LingvoDoc: Working with Cognate Analysis

Natalia Kosheliuk

February 11, 2022

LingvoDoc: working with Cognate Analysis[1]

Kosheliuk Natalia
ORCID 0000-0002-5833-7971
Ivannikov Institute for System Programming of the RAS, Moscow (Russia)
NKoshelyuk@yandex.ru

Abstract. This article offers an overview of one of the main LingvoDoc programs – the Cognate analysis option intended for modern etymological and dialect distance research. It describes in detail how this option works and how to implement it. The advantages of the Cognate Analysis algorithm are also highlighted.
Keywords. LingvoDoc, data mining, linguistics, cognate analysis

1 INTRODUCTION

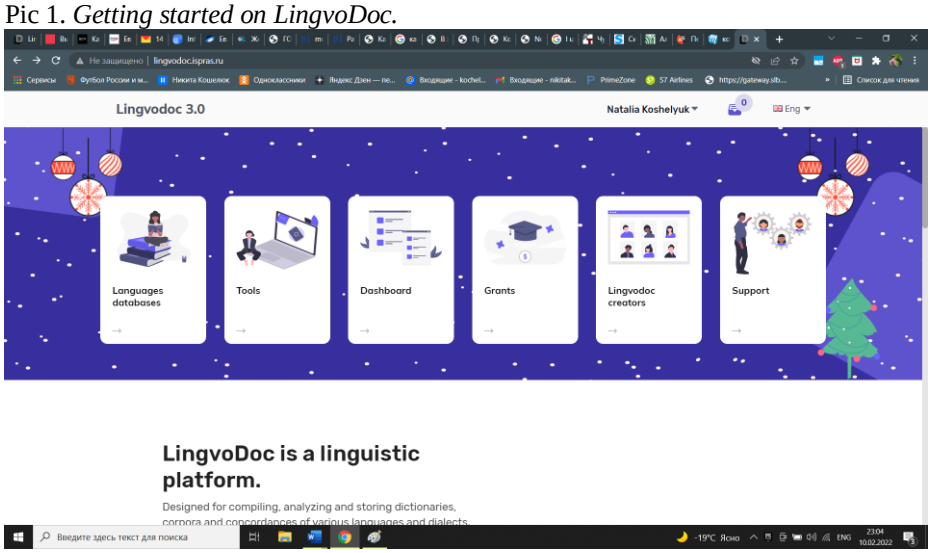
Cognate identification is an important thing for identifying the genetic affinity of languages and their dialects. This method allows linguists to draw conclusions about the development of languages over time and obtain new phonetic and etymological data. Today only a small part of languages has been analyzed in terms of their genetic connections. The main reason is that comparative research in linguistics is still based on the personal work of scientists which is very time-consuming. However, in recent years there has been an active development of methods of computer work on the implementation of these tasks.
In this paper, using the example of the Mansi corpus, we will show how the Cognate Analysis option is implemented in the LingvoDoc and what its advantage is.

2 HOW TO CONDUCT COGNATE ANALYSIS

Initially, work on the LingvoDoc platform begins with standard authorization, downloading the necessary dictionary, or opening an existing one in the database (Pic. 1).

[1] Supported by Russian Science Foundation, project no. 20- 18-00403 'Digital Description of Uralic Languages on the Basis of Big Data'

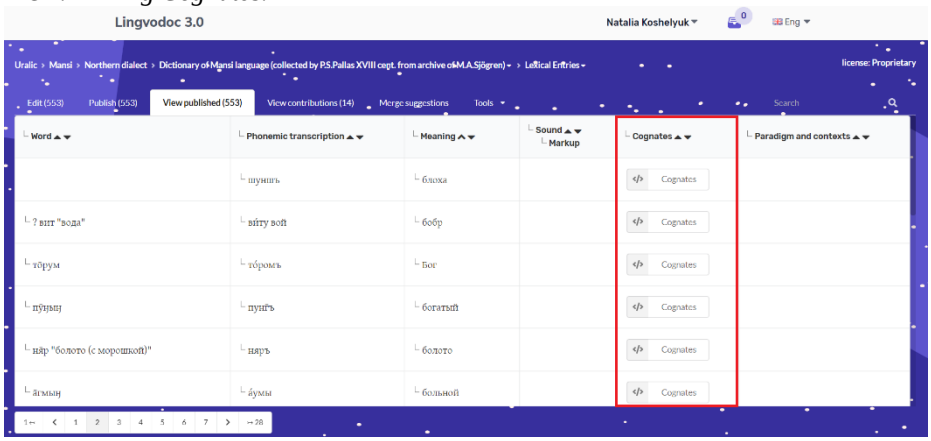
Pic 1. *Getting started on LingvoDoc.*



To obtain the most accurate and reliable result of etymological analysis, if possible, it is necessary to choose the most qualitative and maximally filled dictionaries (Pic. 2).

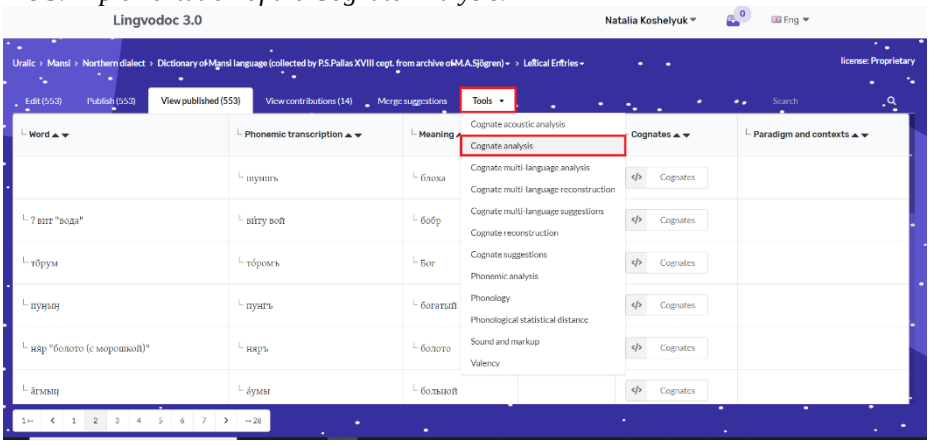
Currently, 23 Mansi dictionaries are connected by etymologies on the Linguistics platform. Most of them are archival texts. The decision to involve them in the analysis is due to the rather high accuracy of the reflection of phonetic oppositions in them. Also, these data are important as one of the first written sources on the Mansi dialects and according to them, you can trace how they have changed over the past 100-250 years.

Pic 2. *Linking Cognates.*



After selecting the necessary linguistic material for Cognate Analysis, you need to go to the "Tools" tab and start the process (Pic. 3).

Pic 3. *Implementation of the Cognate Analysis.*

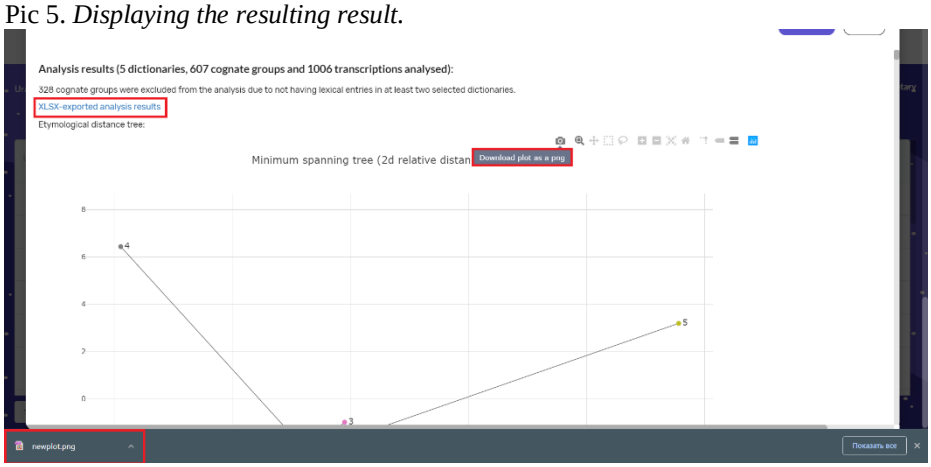


The LingvoDoc platform user also can choose the necessary dictionaries (Pic. 4): you can stop at all available data or limit yourself to at least two dictionaries. Based on our experience, the optimal number of dictionaries subject to processing should not exceed 20. This will be explained by the need to make some editorial changes to combine the rows: manually transferring to reliable – doubtful rows, and to doubtful – single ones. Editorial work with lists of words represented by a large number of columns resulting from the implementation of the Cognate Analysis option is a time-consuming task and can lead to missing important matches.

Pic 4. *Selection of dictionaries necessary for research.*

After starting the process of Cognate Analysis, which takes from 1 to 5 minutes, the user receives a result that reflects the genealogical distance between the sources involved in the analysis, as well as fully painted rows of reliable and unreliable correspondences of vowels and consonants (Pic. 5-6).

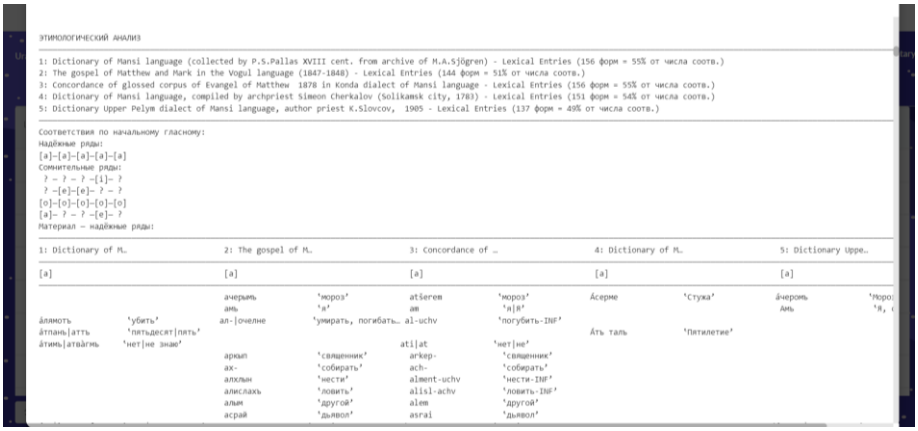
Pic 5. *Displaying the resulting result.*



Briefly outline the principles of the etymological analysis algorithm on the Linguistics platform: at the first stage, for each character from the transcription proposed in each particular dictionary, the algorithm calculates its correspondences in words from other dialects of the same language associated with etymologies with this dictionary. The main calculation is carried out by the roots of words, based on the fact that in them

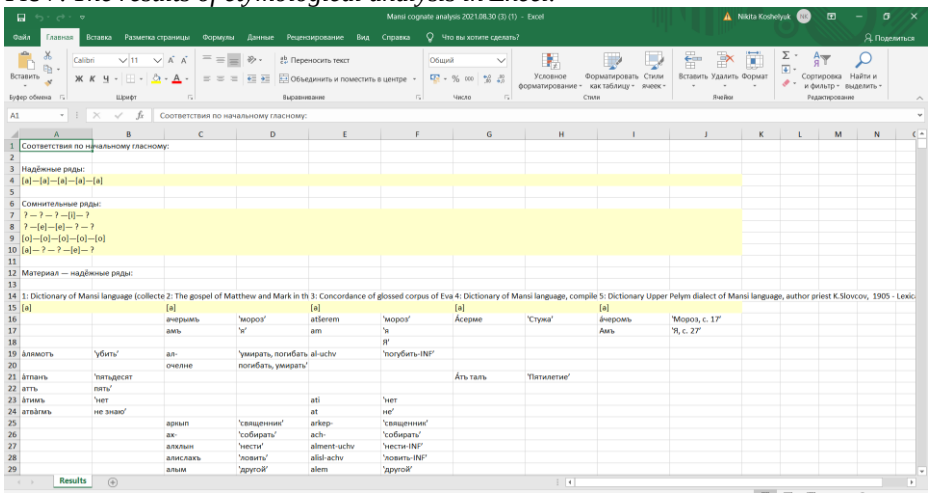
the first vowel corresponds to the first vowel, the first consonant corresponds to the first consonant, the second consonant corresponds to the second. Combinations of vowels and vowels, consonants and consonants are possible. At the output, we get a list of correspondences for each pair of idioms.

Pic 6. *Displaying the resulting result.*



The author of the dictionary has the opportunity to download the results of the analysis in Excel format (Pic. 7), analyze, verify the correctness of transcriptions and etymologies that led to non-standard lines of correspondence, and make adjustments to transcription and etymology. Further, the above algorithm can be restarted again already on the material verified by the author.

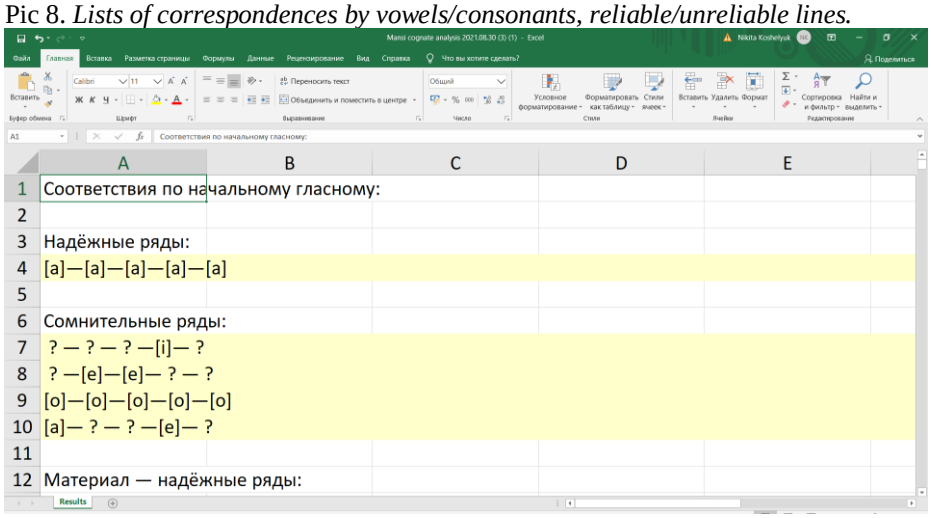
Pic 7. *The results of etymological analysis in Excel.*



	A	B	C	D	E	F	G	H	I	J
1	Соответствия по начальному гласному:									
2										
3	Надёжные ряды:									
4	[a]—[a]—[a]—[a]—[a]									
5										
6	Сомнительные ряды:									
7	? — ? — ? —[i]— ?									
8	? —[e]—[e]— ? — ?									
9	[o]—[o]—[o]—[o]—[o]									
10	[a]— ? — ? —[e]— ?									
11										
12	Материал — надёжные ряды:									
13										
14	1: Dictionary of Mansi language (collecte	2: The gospel of Matthew and Mark in th	3: Concordance of glossed corpus of Eva	4: Dictionary of Mansi language, compile	5: Dictionary Upper Pelym dialect of Mansi language, author priest K.Slovcov, 1905 - Lexic					
15	[a]	[a]	[a]	[a]	[a]					
16		ачерымь	'мороз'	atŝerem	'мороз'	Ăсерме	'Стужа'	ачеромь	'Мороз, с. 17'	
17		амь	'я'	am	'я			Амь	'Я, с. 27'	
18					Я'					
19	ăлямоть	'убить'	ал-	'умирать, погибать	al-uchv	'погубить-INF'				
20			очелне	погибать, умирать'						
21	ăтпань	'пятьдесят					Ăть таль	'Пятилетие'		
22	ăтть	пять'								
23	ăтимь	'нет			ati	'нет				
24	ăтвăлмь	не знаю'			at	не'				
25			арныл	'священник'	arkep-	'священник'				
26			ак-	'собирать'	ach-	'собирать'				
27			алкыын	'нести'	alment-uchv	'нести-INF'				
28			алислахъ	'ловить'	alisl-achv	'ловить-INF'				
29			алым	'другой'	alem	'другой'				

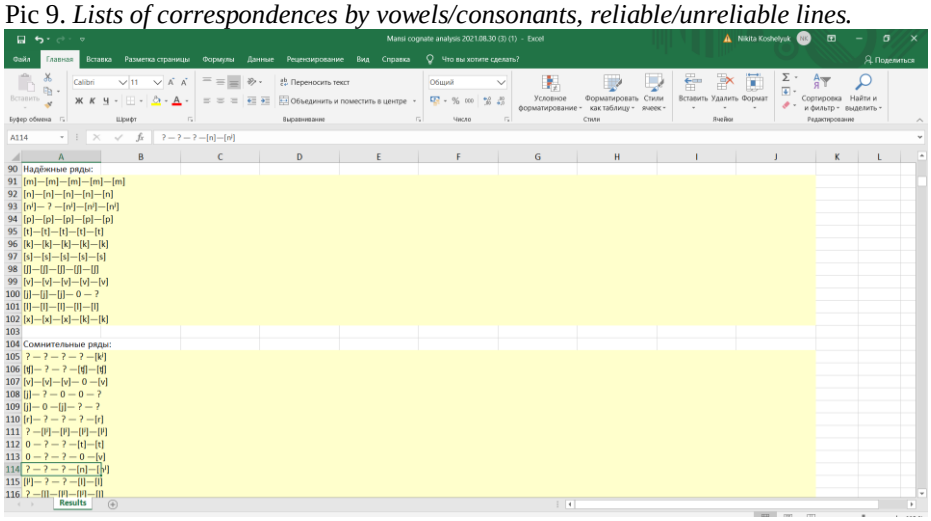
The tool, comparing dictionary materials by a number of parameters (initial vowel, initial consonant, vowel of the first syllable, consonant after the vowel of the first syllable, vowel of the second syllable), finds reliable, doubtful and single series of correlations. The Cognate Analysis option presents the results of comparing dialect materials in the form of a table with detailed lists of correspondences (Pic. 8-9).

Pic 8. *Lists of correspondences by vowels/consonants, reliable/unreliable lines.*



	A	B	C	D	E
1	Соответствия по начальному гласному:				
2					
3	Надёжные ряды:				
4	[a]—[a]—[a]—[a]—[a]				
5					
6	Сомнительные ряды:				
7	? — ? — ? —[i]— ?				
8	? —[e]—[e]— ? — ?				
9	[o]—[o]—[o]—[o]—[o]				
10	[a]— ? — ? —[e]— ?				
11					
12	Материал — надёжные ряды:				

Pic 9. *Lists of correspondences by vowels/consonants, reliable/unreliable lines.*



	A	B	C	D	E	F	G	H	I	J	K	L
90	Надёжные ряды:											
91	[m]—[m]—[m]—[m]—[m]											
92	[n]—[n]—[n]—[n]—[n]											
93	[ṅ]— ? —[ṅ]—[ṅ]—[ṅ]											
94	[p]—[p]—[p]—[p]—[p]											
95	[t]—[t]—[t]—[t]—[t]											
96	[k]—[k]—[k]—[k]—[k]											
97	[s]—[s]—[s]—[s]—[s]											
98	[ʃ]—[ʃ]—[ʃ]—[ʃ]—[ʃ]											
99	[v]—[v]—[v]—[v]—[v]											
100	[j]—[j]—[j]— 0 — ?											
101	[l]—[l]—[l]—[l]—[l]											
102	[x]—[x]—[x]—[k]—[k]											
103												
104	Сомнительные ряды:											
105	? — ? — ? — ? —[k']											
106	[tʃ]— ? — ? —[tʃ]—[tʃ]											
107	[v]—[v]—[v]— 0 —[v]											
108	[j]— ? — 0 — 0 — ?											
109	[j]— 0 —[j]— ? — ?											
110	[r]— ? — ? — ? —[r]											
111	? —[l']—[l']—[l']—[l']											
112	0 — ? — ? —[t]—[t]											
113	0 — ? — ? — 0 —[v]											
114	? — ? — ? —[n]—[ṅ']											
115	[l']— ? — ? —[l]—[l]											
116	? —[l]—[l']—[l']—[l]											

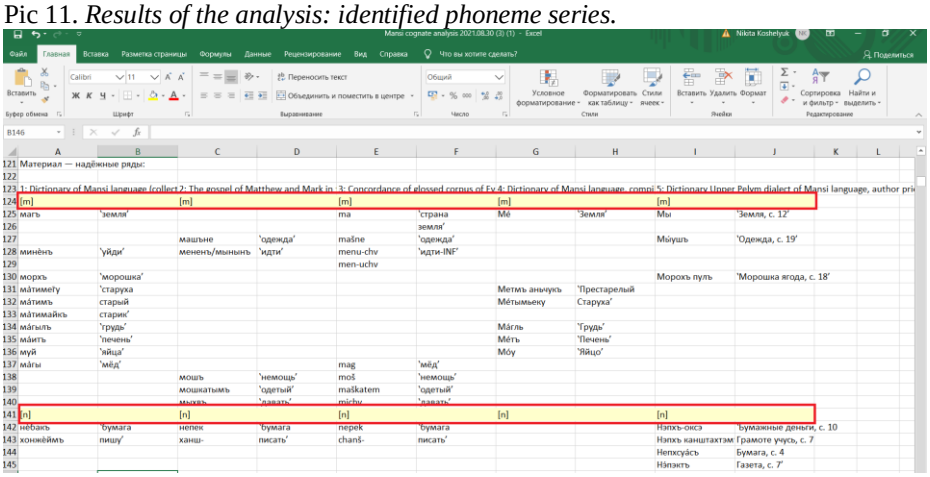
Results

Similar to the analysis presented on the LingvoDoc platform, the Excel table displays the names of dictionaries involved in the study, specific examples of words with their translation, as well as identified consonants and vocalisms (Pic. 10-11). According to the fullness of a particular dictionary, it is likely that the file will not be filled with etymologies for all lines.

Pic 10. *Results of the analysis: dictionary names and identified etymologies.*

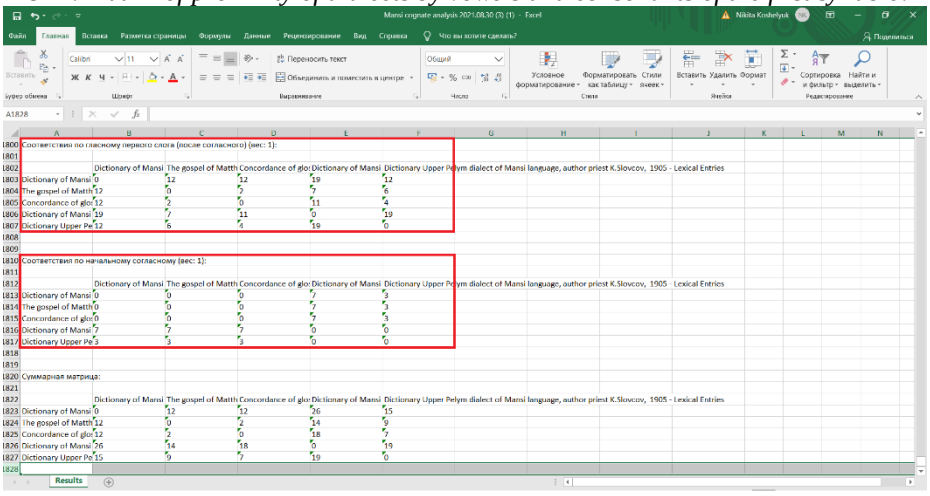
11									
12	Материал — надёжные ряды:								
13									
14	1: Dictionary of Mansi language (collect	2: The gospel of Matthew and Mark in t	3: Concordance of glossed corpus of Ev	4: Dictionary of Mansi language, compil		5: Dictionary Upper Pelym dialect of Mansi la			
15	[a]	[a]	[a]	[a]		[a]			
16		ачерымъ	'мороз'	atŝerem	'мороз'	Асерме	'Стужа'	áчеромъ	'Мороз, с. 17'
17		амъ	'я'	am	'я'			Амъ	'Я, с. 27'
18					'Я'				
19	алямотъ	'убить'	ал-	'умирать, погибать	al-uchv	'погубить-INF'			
20			очелне	погибать, умирать'					
21	ятпанъ	'пятьдесят				Ать таль	'Пятилетие'		
22	ятть	пять'							
23	ятимъ	'нет			ati	'нет			
24	ятвайгмъ	не знаю'			at	не'			
25			аркып	'священник'	arkep-	'священник'			
26			ах-	'собирать'	ach-	'собирать'			
27			алхлын	'нести'	alment-uchv	'нести-INF'			
28			алислахъ	'ловить'	alisl-achv	'ловить-INF'			
29			алым	'другой'	alem	'другой'			
30			асрай	'дьявол'	asrai	'дьявол'			

Pic 11. *Results of the analysis: identified phoneme series.*



Based on the analysis carried out to clarify phonetic and phonological transcription and to calculate regular correlations in words of related languages related by etymologies, at the next stage a program is carried out to determine the proximity of dialects: first, the series of multiple correspondences in the selected group of dictionaries are identified by the vowels of the first syllable and consonants of the first syllable (Pic. 12).

Pic 12. *Matrix of proximity of dialects by vowels and consonants of the first syllable*.



According to Cognate Analysis, based on the identified series of correspondences, all dialects are compared in pairs, the number of transitions

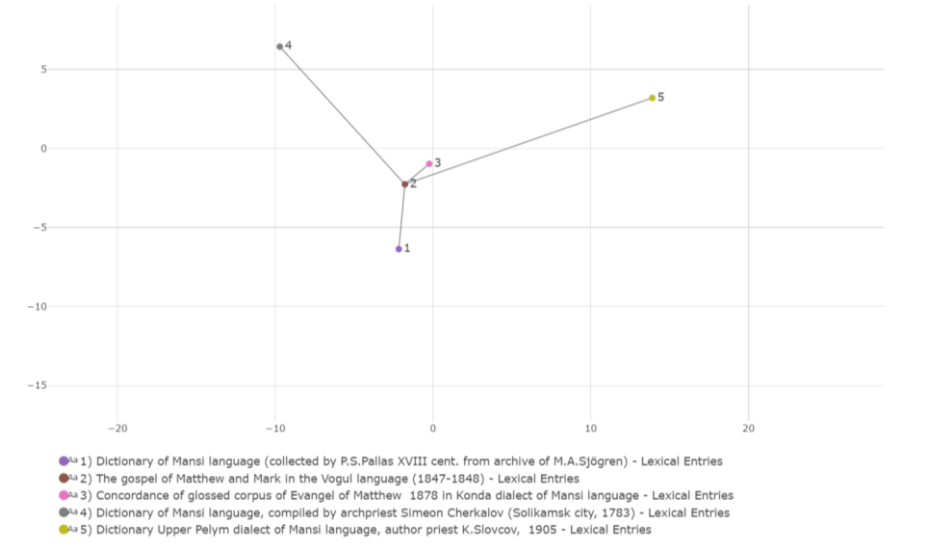
between them is calculated, i.e. a matrix of the analyzed phenomenon is formed: 1) vowels of the first syllable, 2) consonants of the first syllable, 3) sum (Pic. 13).

Pic 13. *The final matrix of dialect proximity*.

1819					
1820	Суммарная матрица:				
1821					
1822		Dictionary of Mansi	The gospel of Matth	Concordance of glo	Dictionary of Mansi	Dictionary Upper Pelym dialect of Mansi language, author priest K.Slovcov, 1905 - Lexical Entries
1823	Dictionary of Mansi	0	12	12	26	15
1824	The gospel of Matth	12	0	2	14	9
1825	Concordance of glo	12	2	0	18	7
1826	Dictionary of Mansi	26	14	18	0	19
1827	Dictionary Upper Pe	15	9	7	19	0
1828						

The correlation of reliable series of the initial position of a word allows the program to calculate the distance between the dialects represented in dictionaries and distribute them into groups. As a result, an etymological and phonetic 3D-model of the distance between the dialects of the selected dictionaries is constructed. The resulting distance represents the number of similar and different phonetic transitions between languages and dialects (Pic. 14).

Pic 14. *3D-model of the distance between dialects.*



As a result, we get a semblance of a family tree, where the dictionaries involved in the analysis are marked with dots with the appropriate color.

CONCLUSION

As can be seen from the proposed review of the implementation of the Cognate Analysis, this program allows you to process large data arrays in semi-automatic mode to establish etymological and phonological systems of dictionaries presented on LingoDoc. The advantage of working with this option is the ability to quickly and fairly reliably refine and recheck the boundaries of many phonetic phenomena, identify special features and processes that have occurred in languages over a given period of time, and so on.

REFERENCES

1. LingvoDoc Homepage, http://lingvodoc.ispras.ru/. Last accessed 02/06/2022