

LingvoDoc: Working with Cognate Analysis

Natalia Kosheliuk

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

February 11, 2022

LingvoDoc: working with Cognate Analysis¹

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. <i>I</i>	Linking	Cognates.
-----------------	---------	-----------

Lingvodoc 3.0	Lingvodoc 3.0										
Uralic > Mansi > Northern dialect > Dictionary of Ma	nsi language (collected by P.S.Pallas XVIII cent	. from archive of MASjögren) - •	 Leftical Entries - 	• •	license: Proprietary						
Edit (553) Publish (553) View published	553) View contributions (14) Mer	ge suggestions Tools •		<u></u>	Search Q						
└ Word ▲ ▼	Phonemic transcription 🔺 🔻	└ Meaning ∧ ▼	└ Sound ▲ ▼ └ Markup	L Cognates 🛦 🗸	└─ Paradigm and contexts ▲ ▼						
	∟ шуншъ	∟ блоха		♦ Cognates							
L-? вит "вода"	∟ ви́ту вой	∟ бобр		Cognates							
∟ _{тõрум}	∟ то́ромъ	L Bor		Cognates							
∟ пўнын	_ пунгъ	∟ богатый		Cognates							
^L ня́р "болото (с морошкой)"	∟ няръ	∟ болото		Cognates							
^L агмын	∟а́умы	^L больной		Cognates							
1⇔ ≤ 1 2 3 4 5 6 7	> → 28			•	. •						

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.

Lingvodoc 3.0			Nata	alia Koshelyuk 🔻 🛛 🧧	as Eng 👻
Uralic > Mansi > Northern dialect > Dictionary of Man	si language (collected by P.S.Pallas XVIII cept. f	rom archive of M	LA.Sjögren) - > LeRical Entries -	• •	license: Proprietary
Edit (553) Publish (553) View published (5	53) View contributions (14) Merg	suggestions	Tools •	·	Search Q
Word	Phonemic transcription	L Meaning	Cognate acoustic analysis	Cognates	Paradigm and contexts
			Cognate analysis		
	∟ шуншгь	∟ блоха	Cognate multi-language analysis Cognate multi-language reconstruction	♦ Cognates	
∟? вит "вода"	∟ ви́ту вой	∟ бобр	Cognate multi-language suggestions Cognate reconstruction	♦ Cognates	
∟тõрум	└ то́ромљ	L Bor	Cognate suggestions Phonemic analysis	Cognates	
∟ пунын	∟ пунгъ	∟ богатый	Phonology Phonological statistical distance	Cognates	
∟ няр "болото (с морошкой)"	⊢няръ	∟ болото	Sound and markup Valency	Cognates	
∟агмың	∟ а́умы	∟ больной		♦ Cognates	
1- < 1 2 3 4 5 6 7 >	⇒28				

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).



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.



ЭТИМОЛОГИЧЕСКИЙ 1: Dictionary of 2: The gospel of 3: Concordance o	АНАЛИЗ Mansi language (coll Matthew and Mark in f glossed corpus of E	ected by P.S.Palla the Vogul language vangel of Matthew	as XVIII cent. from ar a (1847-1848) - Lexici 1878 in Konda diale	rchive of M.A.Sjögr al Entries (144 фор ct of Mansi languag	en) - Lexical Entries м = 51% от числа соот ie - Lexical Entries ((156 форм = 55% о в.) 156 форм = 55% от	п числа соотв.) числа соотв.)		
4: Dictionary or 5: Dictionary Up 	Mansi language, comp per Pelym dialect of i начальному гласному: [a] ы: . ? . ? [o] . ?	lled by arcnpriest Mansi language, au	: Simeon Cherkaiov (St	ollkamsk city, izwa) - Lexical entries (Entries (137 φορω = 4	151 форм = зал от 2% от числа соотв.	числа соотв.;)		
1: Dictionary of	R.	2: The gospel c	of M.	3: Concordance	of _	4: Dictionary	of M.	5: Dictionary	Uppe
[a]		[a]		[a]		[a]		[a]	
Samora.	Subary.	ачерыть амь	*мороз* «д*	atšeren am	*мороз* *я я*	Асерме	¹ Стужа	а́черомь Амь	'Моро: 'Я, с
а́тпань[атть а́тинь]атва́гнь	уомтвдесят пять* Чпятыдесят пять* Чнет не знаю*	аркыл ах- алхлын алислахь алислахь	<pre>*священник* *собирать* *нести* *ловить* *другой*</pre>	ati at arkep- ach- alment-uchv alisl-achv alem	"мет]не" "свлщенник" "собирать" "мести-INF" "ловить-INF" "другой"	А́ть таль	^ч Пятилетие [*]		

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.

										Nikita Koshelyuk	NK 🖸	
Файл Главная	Вставка Размет											Я. Поделитьс
Вставить 🖌	alibri ∨ 11 K K <u>Y</u> + ⊡ + Шрн¢т		= = ≫ · \$\$ ⊓ = = € 至 ⊡ 0	ереносить текст бъединить и помести шнязане	ить в центре 🔹 🧐	-% ∞ % ** ₩κτο Γ	Условное форматирование -	Форматировать Стиль как таблицу - ячеек Стили	Вставить Удалить	Формат	Ат Сортировка н и фильтр - вы Редактировани	Райти и целить *
A1 -	X y fr	Соответствия п	о начальному гласно	wy:								
A 1	В	c	D	F	F	6	н	1	1	ĸ	M	N
1 Соответствия г 2 3 3 Надёжные ряд 4 (a)-(a)-(a)-(a)-(a)-(a)-(a)-(a)-(a)-(a)-	ю начальному глас bi: a)—(a) pagas: - ? - ? (o)—(o)]— ? адёжные рады:	HOMY:										
4 1: Dictionary of 5 [a]	Mansi language (col	lecte 2: The gospel o	Matthew and Mark I	fal	of glossed corpus of	fal	Mansi language, cor	fal	er Pelym dialect of Ma	insi language, aut	for priest K.Slov	cov, 1905 - Lex
16		ачерымъ	'mopos'	atšerem	'mopos'	Асерме	'Стужа'	а́черомъ	'Mopos, c. 17'			
7		амъ	'8'	am	' 8			Амъ	'Я, с. 27'			
.8 .9 а́лямоть 20	'убить'	ал- очелне	умирать, погиб погибать, умира	ать al-uchv iть'	я' 'погубить-INF'							
1 атпанъ	'пятьдесят					А́ть таль	'Пятилетие'					
A ATMAN	Suer			ati	Suer							
4 атрагиъ	не знаю'			at	He							
5		аркыл	'священник'	arkep-	священник'							
16		ax-	'собирать'	ach-	'собирать'							
7		алхлын	'нести'	alment-uchy	'HECTH-INF'							
8		алислахь	'ловить'	alisl-achv	'noswite-INF'							
9		алым	'apyro#'	alem	'apyro#'							
Base	lte O											

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 correspond	lences by v	owels/consond	ints, reliable/un	reliable lines.	
- E •	5 • 0 • 0		Mansi cognate analysis 2021.08.30 (3) (1) - Exo	el	🔺 Nikita Koshelyuk 🛞 🖬 —	o x
Файл	Главная Вставка Разметка страницы Формулы Данные	Рецензирование Вид Сп	правка 🛛 Что вы хотите сделать?			Я. Поделиться
Вставить Буфер обе	$ \begin{array}{c} \overset{X}{\underset{{}_{\scriptstyle{\scriptstyle{\scriptstyle{\scriptstyle{\scriptstyle{\scriptstyle{\scriptstyle{\scriptstyle{\scriptstyle\scriptstyle{\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle\scriptstyle$	ар Переносить текст Объединить и поместить в це Выравнивание	нитре - Сбщияй	словное пирование - кактаблицу - ячеек- Слин	ть Удалить Формат Выйке Родицания Выйке	Окти и заить -
AL	 з соответствия по начальному гл 	B	C	D	F	*
1	Соответствия по начальном	у гласному:	6	0	_	
2						
3	Надёжные ряды:					
4	[a]—[a]—[a]—[a]—[a]					
5						
6	Сомнительные ряды:					
7	? — ? — ? —[i]— ?					
8	? —[e]—[e]— ? — ?					
9	[0]—[0]—[0]—[0]					
10	[a]— ? — ? —[e]— ?					
11						
12	Материал — надёжные ря	ды:				
	Results (+)			1.4		Þ

D' O Itata of oom ndancas hu vowals/consonan raliable/uproliable lin



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

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.

материал —	надежные ряды:								
1: Dictionary o	of Mansi language (co	ellect 2: The gospel of	of Matthew and Mark	in 13: Concordance	of glossed corpus of	Ev 4: Dictionary of	of Mansi language, co	mpil5: Dictionary U	Jpper Pelym dialect o
[a]		[a]		[a]		[a]		[a]	
		ачерымъ	'mopo3'	atšerem	'мороз'	Асерме	'Стужа'	а́черомъ	'Мороз, с. 17 '
		амъ	`я'	am	'я			Амъ	'Я, c. 27'
					Я'				
алямотъ	'убить'	ал-	умирать, поги	бать al-uchv	'погубить-INF'				
		очелне	погибать, умир	ать'					
атпанъ	`пятьдесят					А́ть таль	'Пятилетие'		
атть	пять'								
атимъ	`нет			ati	'нет				
атва̀гмъ	не знаю'			at	не'				
		аркып	'священник'	arkep-	'священник'				
		ax-	'собирать'	ach-	'собирать'				
		алхлын	'нести'	alment-uchv	'нести-INF'				
		алислахъ	'ловить'	alisl-achv	'ловить-INF'				
		алым	'другой'	alem	'другой'				
		асрай	'дьявол'	asrai	'дьявол'				

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

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

≂ ⊸ - د+ ∎				Mansi o	ognate analysis 2021.08.3	30 (3) (1) - Excel			Nikita Koshelyuk 📧	- 19	- 0	
Файл Главная Е	ставка Разметка стра	ницы Формулы Да	нные Рецензирован	ие Вид Справка							. Поделить	
Вставить 🖌 Санк	n ∨11 ∨ A K <u>H</u> + ⊞ + <u>∆</u> + ,		- 👌 Переносить 📰 🔯 Объединить	текст и поместить в центре	. 😨 - % 000 💃	У Условное форматирован	Форматировать С ие - как таблицу - як	Тили Вставить Удалит	Бормат	АТ Я пировка На фильтр - выр	О или и елить *	
Буфер объека ГС	Шрифт		Быравнивание		Fy Mincho		Стили	Prestan	Pe	дактирование		
B146 * i	$\times \checkmark f_x$											¥
A	В	c	D	E	F	G	н	1	1	к	L L	
121 Материал — на	дёжные ряды:											
122												1
123 1: Dictionary of I	Aansi language (colle	ect 2: The gospel of M	atthew and Mark in	3: Concordance o	f glossed corpus of t	Ev 4: Dictionary of M	lansi language, com	ni 5: Dictionary Upper	Pelym dialect of M	ansi langua	ge, author p	orie
124 (m)		(m)		(m)		[m]		(m)				
125 магь	'земля'			ma	'страна	Mé	'Земля'	Мы	'Земля, с. 12'	1		
126					земля'							
127		машъне	`одежда'	mašne	'одежда'			Мы́ушъ	'Одежда, с. 19'			
128 минёнъ	'уйди'	мененъ/мынынъ	`идти′	menu-chv	'идти-INF'							
129				men-uchv								
130 морхъ	`морошка'							Морохъ пуль	'Морошка ягода, с	. 18'		
131 матимету	`старуха					Метмъ аньчукъ	'Престарелый					
132 ма́тимъ	старый					Метымьеку	Старуха'					
133 матимайкъ	старик'											
134 ма́гылъ	'грудь'					Магль	'Грудь'					
135 ма́ить	'печень'					Меть	'Печень'					
136 мүй	'яйца'					Móy	'Яйцо'					
137 ма́гы	`мёд'			mag	`мёд'							
138		мошъ	`немощь'	moš	'немощь'							
139		мошкатымъ	`одетый'	maškatem	`одетый'							
140		MARYER	'nasats'	michy	'павать'							
141 [n]		[n]		[n]		[n]		[n]				
142 не́бакъ	бумага	непек	ъумага	nepek	бумага			Нэпхъ-оксэ	Ъумажные деньги	, c. 10		
143 хонжёймъ	пишу′	ханш-	писать'	chanš-	писать'			Нэпхъ канштахтэм	Грамоте учусь, с. 7	(
144								Непхсуа́съ	Бумага, с. 4			
145								Нэ́пэкть	Газета, с. 7			

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).

									Δ •	likita Koshelyuk	· · · · · · · · · · · · · · · · · · ·	
айл Главная Вст		ицы Формулы		рование Вид		Q Что вы хотите сдея						Я. Поделить
Calibri	v 11 v A <u>4</u> • ⊞ • <u>à</u> •		 	осить текст инить и поместить	ицентра -	Общий V	Условное форматирование	Форматировать Стили • как таблицу • ячеек •	Вставить Удалить	тормат €	Сортировка Н и фильтр - вы	р Іайти и целить *
увер обмена 🕠	Workpr	<u>a</u>	Выракиия	9-974		Hirona	il .	Стити	Stellor		Редактирозание	
1828 - 1 2	√ fr											
Δ	В	с	D	E	F	G	н			К	L M	N
00 Соответствия по гл 101	асному первого слог	а (после согласно	ro) (sec: 1):									
02	Dictionary of Mansi T	he gospel of Matth	Concordance of glos	Dictionary of Mar	si Dictionary	Upper Pelym dialect of	Mansi language, aut	hor priest K.Slovcov, 1905	- Lexical Entries			
03 Dictionary of Mansi	0 1	2	12	19	12							
04 The gospel of Matth	12 0		2	7	6							
05 Concordance of glo:	12 2		0	11	4							
ID6 Dictionary of Mansi	19 7		11	0	19							
07 Dictionary Upper Pe	12 6		4	19	0							
08												
109												
10 Соответствия по на	чальному согласном	vy (sec: 1):										
11												
12	Dictionary of Mansi T	he gospel of Matth	Concordance of glos	Dictionary of Mar	isi Dictionary	Upper Peyrn dialect of	Mansi language, aut	hor priest K.Slovcov, 1905	 Lexical Entries 			
13 Dictionary of Mansi	0 0		0	7	3							
14 The gospel of Matth	0 0		0	7	3							
15 Concordance of glos	0 0		0	7	3							
16 Dictionary of Mansi	7 7		7	0	0							
17 Dictionary Upper Pe	3 3		3	0	0							
18												
19												
20 Суммарная матриц	a:											
21												
22	Dictionary of Mansi T	he gospel of Matth	Concordance of glos	Dictionary of Mar	si Dictionary	Upper Pelym dialect of	Mansi language, aut	hor priest K.Slovcov, 1905	- Lexical Entries			
23 Dictionary of Mansi	0 1	2	12	26	15							
24 The gospel of Matth	12 0		2	14	9							
25 Concordance of glo:	12 2		0	18	7							
26 Dictionary of Mansi	26 1	4	18	0	19							
27 Dictionary Upper Pe	15 9		7	19	0							
28												

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 матри	ta:								
1821									
1822	Dictionary of Mansi	The gospel of Matth	Concordance of glos	Dictionary of Mansi	Dictionary Upper Pe	ym dialect of Mansi	language, author prie	est 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