"LingvoDoc" as a Solution to the Problem of Conservation of Endangered Languages (Using Example of the Mansi Language)

Natalia Kosheliuk

July 4, 2022

«LingvoDoc» as a solution to the problem of conservation of endangered languages (using example of the Mansi language)[1]

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

Abstract. This article, using the example of the Russian language, provides an overview of the main problems concerning the endangered languages of Russia, and describes ways to solve them using a linguistic platform LingvoDoc.

Keywords. LingvoDoc, data mining, linguistics, endangered languages, conservation, research.

1 INTRODUCTION

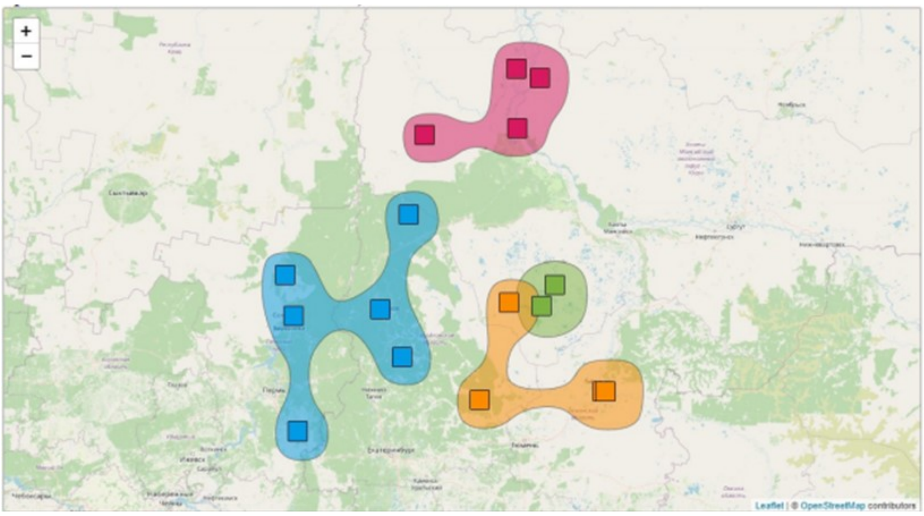
The Mansi language is the language of native speakers living in the vast territory of Northwestern Siberia. Currently, we are witnessing a catastrophic situation: this language is rapidly dying. According to the annual monitoring of the Institute of Ob-Ugric Research and Development, by 2019, about 1,000 people were still alive who were able to speak the North Mansi dialect, the only one preserved today. These are mostly older people.

It is worth noting that the last speakers of western and southern dialects were recorded in the first decades of the XX century, and the latest data on the eastern Yukonda dialect was recorded in 2013 during an expedition to the Khanty-Mansi Autonomous district by an em-

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

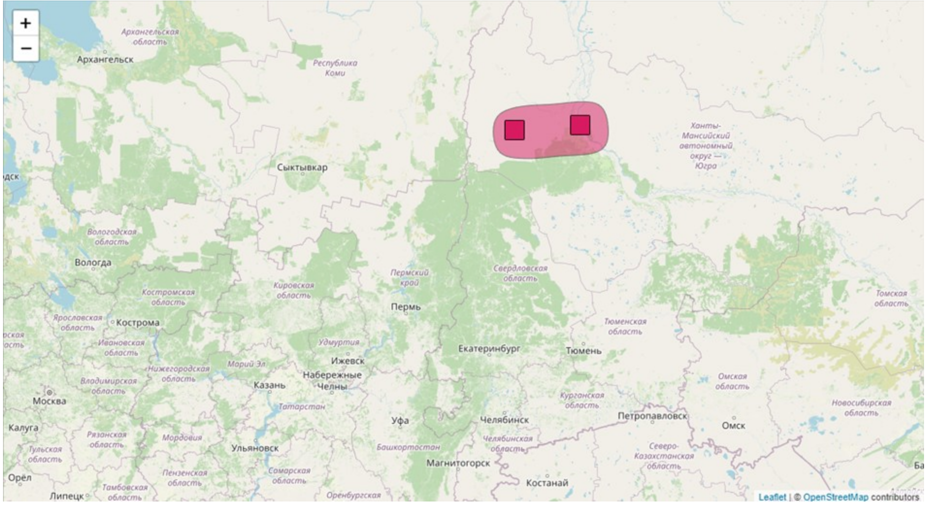
ployee of the Institute of Linguistics of the Russian Academy of Sciences. In the village of Shugur, only 2 native speakers were found, Selivanova Elizaveta Ivanovna (89 years old) and Shivtorov Maxim Semenovich (74 years old). Elizaveta Ivanovna passed away 5 days after the departure of the linguist from Shugur, and Maxim Semenovich – in 2018. Thus, the eastern Mansi dialect disappeared literally before our eyes[2] (Pic. 1-2).

Pic 1. The area of distribution of the Mansi language in the XVIII-XIX cent.



[2] The speech recording of these last two native speakers is now stored on «LingvoDoc». It can be found at the link:: http://lingvodoc.ispras.ru/dictionary/657/3/perspective/657/4/view.

Pic 2. The area of distribution of the Mansi language in the XXI cent.



Another problem concerning the Mansi language, as well as other endangered languages, is the difficulty of conducting scientific expeditions. It often turns out that it is often problematic for researchers to get to the places of residence of native speakers: Mansi places of residence are difficult to access for expeditions and require significant resources – physical and financial investments. It is not uncommon for a situation when upon arrival it turns out that the Mansi of those places do not really speak their own language.

This also raises another important problem concerning many languages of Russia – the limited language corpus and, in general, the inaccessibility of most materials for linguists, historians, ethnographers and other specialists. It is known that at present many archival philological documents over the past decades have been tabulated (in particular, the first Mansi Cyrillic translations, which were digitized in the archives of St. Petersburg in 2018-2019. currently, for this reason, they are no longer available) or lost due to transfer to foreign universities (for example, as happened with the archive of A. P. Dulzon). Another point that needs to be mentioned is that many researchers are afraid to publish their materials, transfer them to electronic form, as there are worries about a possible data leak.

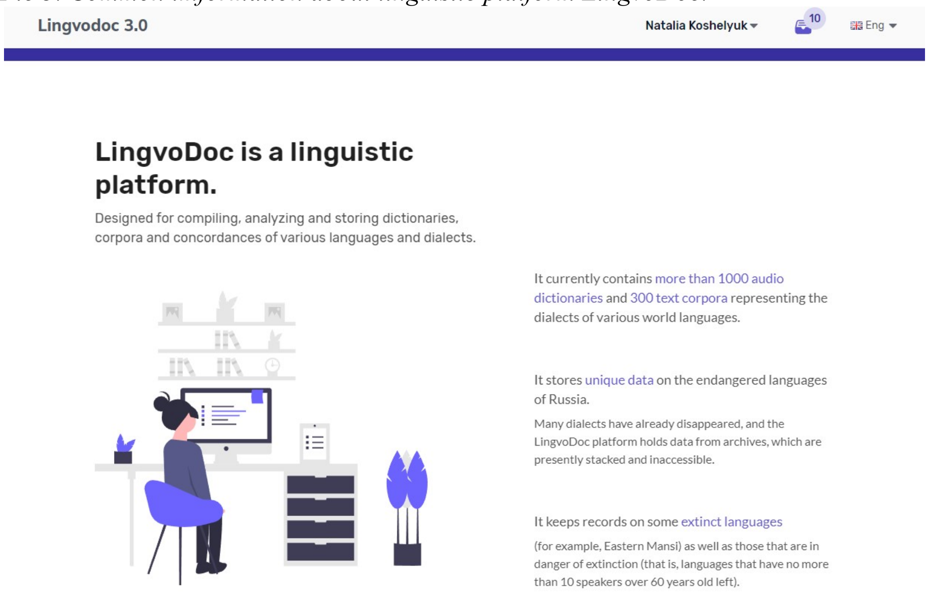
Thus, if no active actions are taken to solve these problems, in the very near future the cultural heritage of Russia in the form of valuable linguistic and archival material may simply disappear.

2 HOW LINGVODOC CAN SOLVE THE PROBLEMS OF ENDANGERED LANGUAGES

What solutions does LingvoDoc offer to contribute to the preservation of endangered languages in Russia:

1. The project management (Yu. V. Normanskaya, A. I. Avetisyan, O. D. Borisenko) has a principled position: placing all data in the public domain. Western resources do not practice this – most platforms provide the researcher with data on request.
2. Anyone can register and post their material. Currently, more than 1.000 dictionaries and text corpora have been uploaded to LingvoDoc (Pic. 3).

Pic 3. Common information about linguistic platform LingvoDoc.



LingvoDoc is a linguistic platform.

Designed for compiling, analyzing and storing dictionaries, corpora and concordances of various languages and dialects.

It currently contains more than 1000 audio dictionaries and 300 text corpora representing the dialects of various world languages.

It stores unique data on the endangered languages of Russia.

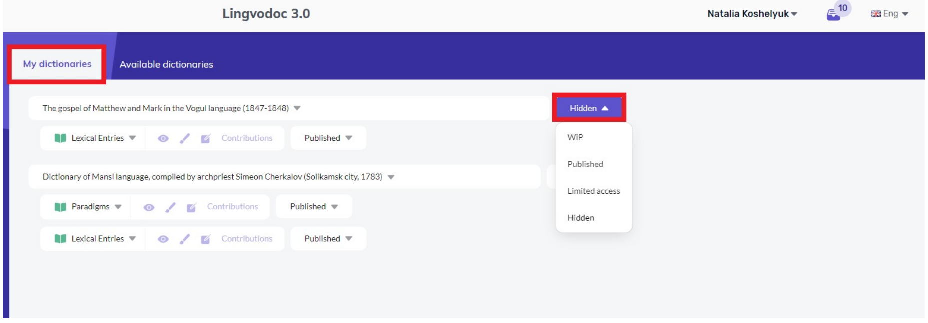
Many dialects have already disappeared, and the LingvoDoc platform holds data from archives, which are presently stacked and inaccessible.

It keeps records on some extinct languages

(for example, Eastern Mansi) as well as those that are in danger of extinction (that is, languages that have no more than 10 speakers over 60 years old left).

3. If the user believes that the material he is working on at the LingvoDoc needs to be improved, it can be hidden in the program settings (My dictionaries -> Hide the dictionary), and published when the author deems it necessary (Pic. 4).

Pic 4. Steps to switch the dictionary to invisible mode.

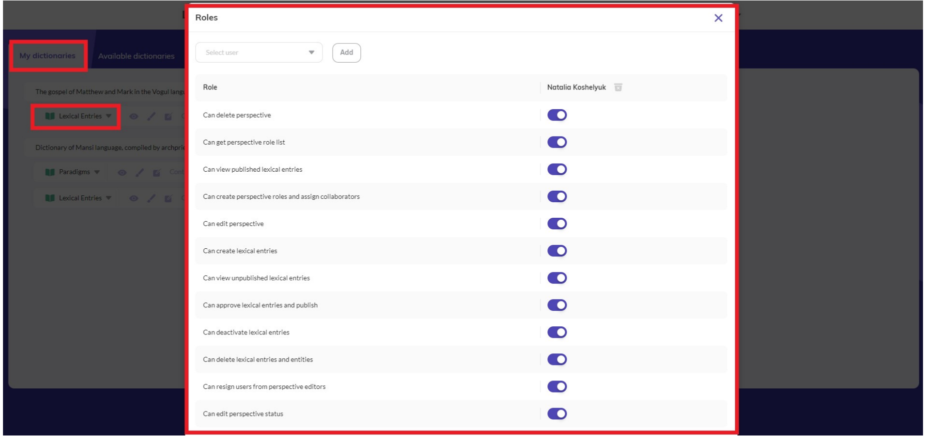


The gospel of Matthew and Mark in the Vogul language (1847-1848)

Dictionary of Mansi language, compiled by archpriest Simeon Cherkalov (Solikamsk city, 1783)

4. To prevent plagiarism (Pic. 5), the platform has implemented the option of granting rights to work and edit LingvoDoc resources: you can adjust the list of people who can interact with your corpus or dictionary (My dictionaries -> Lexical entries -> Roles -> Choose the right one).

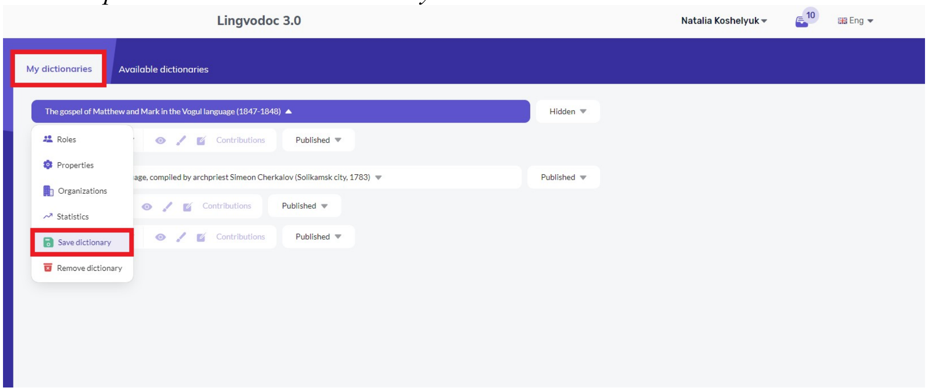
Pic 5. Steps for granting rights to use and edit the dictionary.


My dictionaries
Available dictionaries
The gospel of Matthew and Mark in the Vogul langu
Lexical Entries
Dictionary of Mansi language, compiled by archpri
Paradigms
Lexical Entries
Roles
Select user
Add

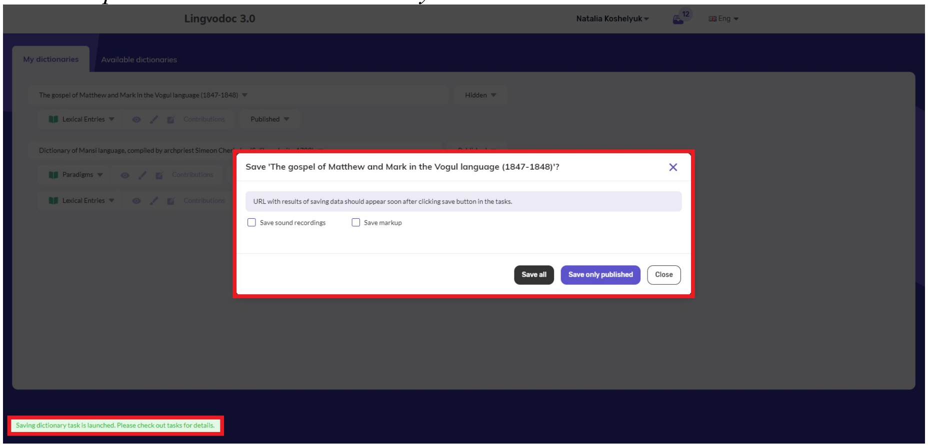
Role	Natalia Koshelyuk
Can delete perspective	●
Can get perspective role list	●
Can view published lexical entries	●
Can create perspective roles and assign collaborators	●
Can edit perspective	●
Can create lexical entries	●
Can view unpublished lexical entries	●
Can approve lexical entries and publish	●
Can deactivate lexical entries	●
Can delete lexical entries and entities	●
Can resign users from perspective editors	●
Can edit perspective status	●

5. Each dictionary or corpus posted on the platform can be registered as IPR on the Gosuslugi website (https://www.gosuslugi.ru – for Russia). This legally secures the copyright to the posted data (Pic. 6-9).

Pic 6. Step 1 to download the dictionary to exe.



Pic 7. Step 2 to download the dictionary to exe.



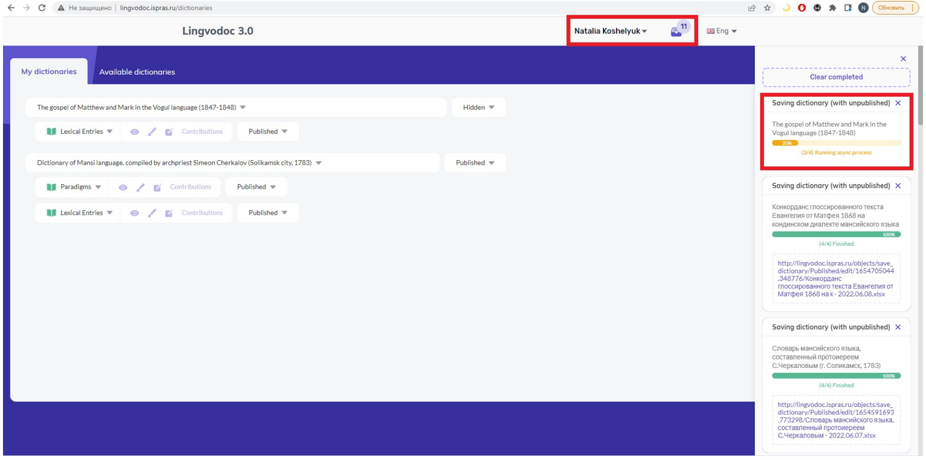
Save 'The gospel of Matthew and Mark in the Vogul language (1847-1848)'?

URL with results of saving data should appear soon after clicking save button in the tasks.

□ Save sound recordings □ Save markup

Saving dictionary task is launched. Please check out tasks for details.

Pic 8. Downloading the dictionary is in progress.



Не защищено | lingvodoc.ispras.ru/dictionaries

Lingvodoc 3.0

Natalia Koshelyuk 11 Eng

My dictionaries | Available dictionaries

The gospel of Matthew and Mark in the Vogul language (1847-1848) — Hidden
- Lexical Entries — Contributions — Published

Dictionary of Mansi language, compiled by archpriest Simeon Cherkalov (Solikamsk city, 1783) — Published
- Paradigms — Contributions — Published
- Lexical Entries — Contributions — Published

Clear completed

Saving dictionary (with unpublished)

The gospel of Matthew and Mark in the Vogul language (1847-1848)

20%

(3/4) Running async process

Saving dictionary (with unpublished)

Конкорданс глоссированного текста Евангелия от Матфея 1868 на кондинском диалекте мансийского языка

100%

(4/4) Finished

http://lingvodoc.ispras.ru/objects/save_dictionary/Published/edit/1654705044.348776/Конкорданс глоссированного текста Евангелия от Матфея 1868 на к - 2022.06.08.xlsx

Saving dictionary (with unpublished)

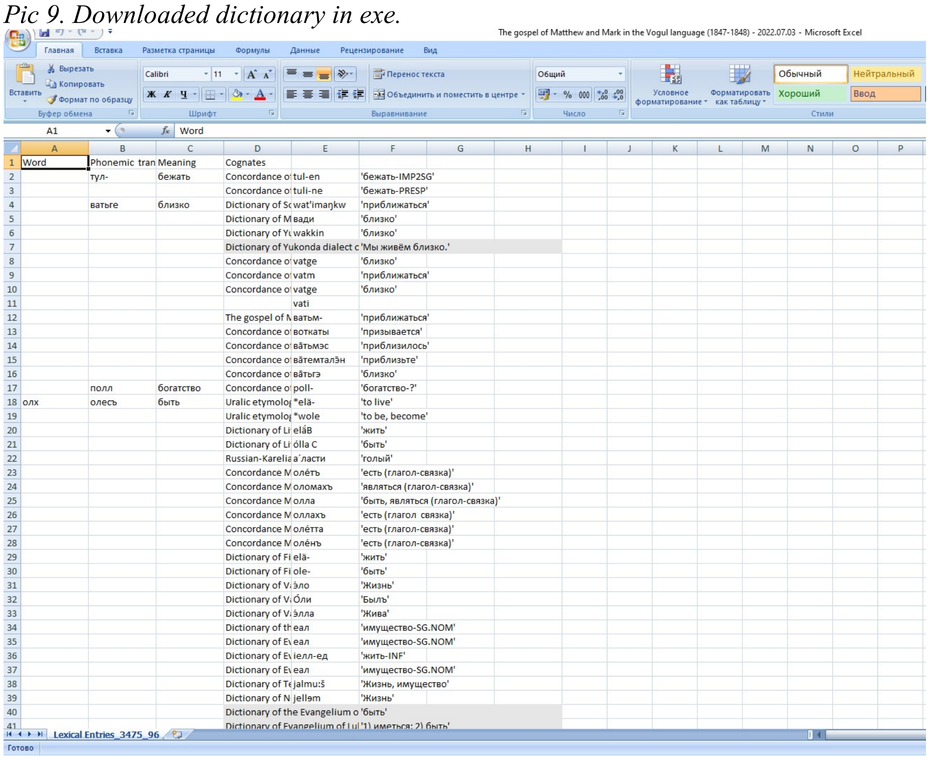
Словарь мансийского языка, составленный протоиереем С.Черкаловым (г. Соликамск, 1783)

100%

(4/4) Finished

http://lingvodoc.ispras.ru/objects/save_dictionary/Published/edit/1654591693.773298/Словарь мансийского языка, составленный протоиереем С.Черкаловым - 2022.06.07.xlsx

Pic 9. Downloaded dictionary in exe.



The gospel of Matthew and Mark in the Vogul language (1847-1848) - 2022.07.03 - Microsoft Excel

A1 — Word

	A	B	C	D	E	F
1	Word	Phonemic tran	Meaning	Cognates		
2		тул-	бежать	Concordance o	tul-en	'бежать-IMP2SG'
3				Concordance o	tuli-ne	'бежать-PRESP'
4		ватьге	близко	Dictionary of Sc	wat'imaŋkw	'приближаться'
5				Dictionary of M	вади	'близко'
6				Dictionary of Yu	wakkin	'близко'
7				Dictionary of Yukonda dialect c		'Мы живём близко.'
8				Concordance o	vatge	'близко'
9				Concordance o	vatm	'приближаться'
10				Concordance o	vatge	'близко'
11					vati	
12				The gospel of M	ватьм-	'приближаться'
13				Concordance o	воткаты	'призывается'
14				Concordance o	вāтьмэс	'приблизилось'
15				Concordance o	вāтемталэ̄н	'приблизьте'
16				Concordance o	вāтьгэ	'близко'
17		полл	богатство	Concordance o	poll-	'богатство-?'
18	олх	олесъ	быть	Uralic etymolog	*elä-	'to live'
19				Uralic etymolog	*wole	'to be, become'
20				Dictionary of Li	eläB	'жить'
21				Dictionary of Li	ólla C	'быть'
22				Russian-Karelia	а´ласти	'голый'
23				Concordance M	олéтъ	'есть (глагол-связка)'
24				Concordance M	оломахъ	'являться (глагол-связка)'
25				Concordance M	олла	'быть, являться (глагол-связка)'
26				Concordance M	оллахъ	'есть (глагол связка)'
27				Concordance M	олéтта	'есть (глагол-связка)'
28				Concordance M	олéнъ	'есть (глагол-связка)'
29				Dictionary of Fi	elä-	'жить'
30				Dictionary of Fi	ole-	'быть'
31				Dictionary of Va	э́ло	'Жизнь'
32				Dictionary of Va	Óли	'Былъ'
33				Dictionary of Va	э́лла	'Жива'
34				Dictionary of th	еал	'имущество-SG.NOM'
35				Dictionary of Ev	еал	'имущество-SG.NOM'
36				Dictionary of Ev	иелл-ед	'жить-INF'
37				Dictionary of Ev	еал	'имущество-SG.NOM'
38				Dictionary of Te	jalmu:š	'Жизнь, имущество'
39				Dictionary of N	jelləm	'Жизнь'
40				Dictionary of the Evangelium o		'быть'
41				Dictionary of Evangelium of Lul		'1) иметься; 2) быть'

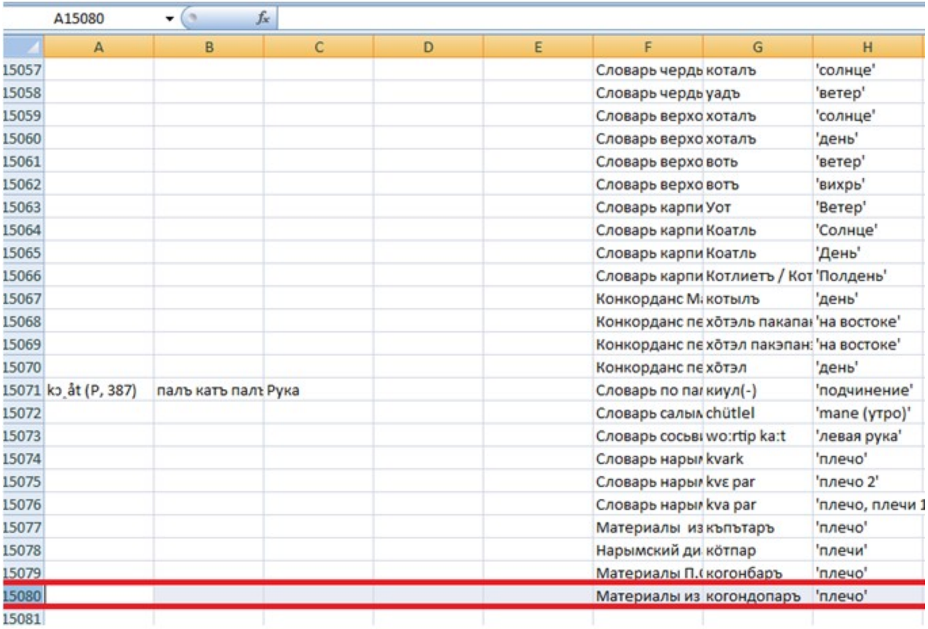
Lexical Entries_3475_96

Готово

It is also necessary to mention another opportunity provided by Linguodoc – the publication of monographs as an indicator of the publication activity of various scientists and scientific institutions.

For today, the amount of data that has already been posted on the LingvoDoc (more than 1,000 sources) allows us to reach the level of publishing monographs from 3-5 books per year for one author. This is ensured by the published amount of data on the LingvoDoc and its processing by a special algorithm that allows you to convert drafts of a future monograph in pdf format: a list of related words, the editing of which can eventually serve as the basis of an etymological dictionary with an average volume of 300-500 pages (Pic. 10).

Pic 10. A draft of the future etymological Mansi dictionary based on one source posted on the platform.



A15080

	A	B	C	D	E	F	G	H
15057						Словарь чердь	коталъ	'солнце'
15058						Словарь чердь	уадъ	'ветер'
15059						Словарь верхо	хоталъ	'солнце'
15060						Словарь верхо	хоталъ	'день'
15061						Словарь верхо	воть	'ветер'
15062						Словарь верхо	вотъ	'вихрь'
15063						Словарь карпи	Уот	'Ветер'
15064						Словарь карпи	Коатль	'Солнце'
15065						Словарь карпи	Коатль	'День'
15066						Словарь карпи	Котлиетъ / Кот	'Полдень'
15067						Конкорданс М	котылъ	'день'
15068						Конкорданс пе	хо̄тэль пакапа	'на востоке'
15069						Конкорданс пе	хо̄тэл пакэпан	'на востоке'
15070						Конкорданс пе	хо̄тэл	'день'
15071	kɔ_åt (P, 387)	палъ катъ палъ	Рука			Словарь по па	киул(-)	'подчинение'
15072						Словарь салым	chütlel	'mane (утро)'
15073						Словарь сосьв	wo:rtip ka:t	'левая рука'
15074						Словарь нарым	kvark	'плечо'
15075						Словарь нарым	kvɛ par	'плечо 2'
15076						Словарь нарым	kva par	'плечо, плечи 1
15077						Материалы из	къпътаръ	'плечо'
15078						Нарымский ди	кӧтпар	'плечи'
15079						Материалы П.(	когонбаръ	'плечо'
15080						Материалы из	когондопаръ	'плечо'
15081								

In the old days, it took linguists decades to carry out such work – collecting material, linking each lexeme with an etymological link and subsequent editorial work on one book. A review of this feature clearly demonstrates how such tasks can now be handled in a few weeks and months.

CONCLUSION

The opportunities provided by the LingvoDoc for the preservation of endangered languages are an example of how each of the linguists can work with their own languages, as well as quickly publish their articles and monographs. And for higher educational institutions and scientific institutes, this is an example of how each of them can significantly increase the indicators of their institutions in terms of publication activity.

REFERENCES

1. LingvoDoc Homepage, http://lingvodoc.ispras.ru/. Last accessed: 24/06/2022