EasyChair Preprint

№ 5387

LingvoDoc: New Opportunities for Graphic and Phonetic Analysis of Endangered Languages

Natalia Koshelyuk

April 25, 2021

LingvoDoc: new opportunities for graphic and phonetic analysis of endangered languages[1]

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

Abstract. The article presents description and user experience on LingvoDoc, software for graphics and phonetics analysis. Advantages of the LingvoDoc platform over previously used methods are revealed based on the archive materials study of the endangered Mansi language.

Keywords. LingvoDoc, data mining, phonetic and graphic analysis, linguistics.

1 INTRODUCTION

LingvoDoc is a multifunctional linguistic platform designed for generation, analysis and storage of linguistic data of various languages and dialects. It was elaborated under the guidance of J.V. Normanskaya in 2012, and today it is a platform for creating dictionaries, corpora, concordances and texts of various languages of the world (the primary goal of the program is to study the disappearing and endangered languages of Russia). Starting from 2017 LingvoDoc has made it possible to conduct a complete analysis of the linguistic data with automated verification of the processed data; the linguists have access to such options as phonetic, phonemic and acoustic researches, they can search for etymologies, reconstruct cognates of dialects and several languages, add sound and markup prepared in the Praat system and much more (more details at http://lingvodoc.ispras.ru/). Today, it is a dynamically developing program based at the Institute of Programming Systems of the Russian

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

Academy of Sciences (Moscow), and it is constantly updating its options and features.

This article represents a step-by-step description for working in LingvoDoc using the options for creating dictionaries, filling and editing them, adding parallels and etymological connections necessary for conducting a graphic and phonetic analysis of archival linguistic data.

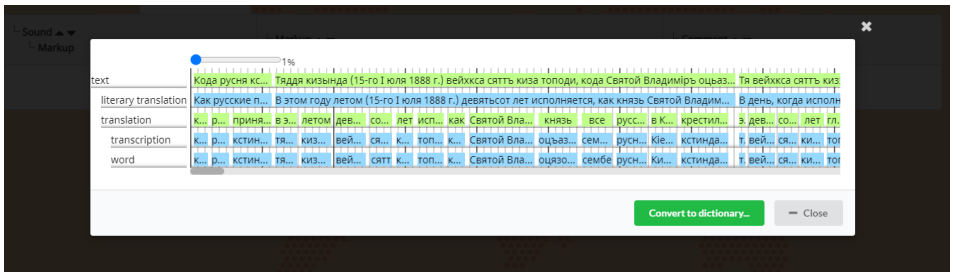
2 HOW TO ANALYSE ARCHIVAL LINGUISTIC DATA

Work with the sources on LingvoDoc platform always starts either with registering in the system or logging into a personal account. After this step, the user acquires access to the entire linguistic database of the platform (in some cases, with prior permission of the author of a particular dictionary or corpus) and all the options, including creating your own dictionary.

Archival data discovered by the researchers or field data collected during expeditions often become both the basis for a future electronic dictionary and the data on which further research is based. In our case, an example of using the LingvoDoc platform for conducting a modern linguistic research is the archive of the Western Mansi dialect of Father Konstantin Slovtsov (1905).

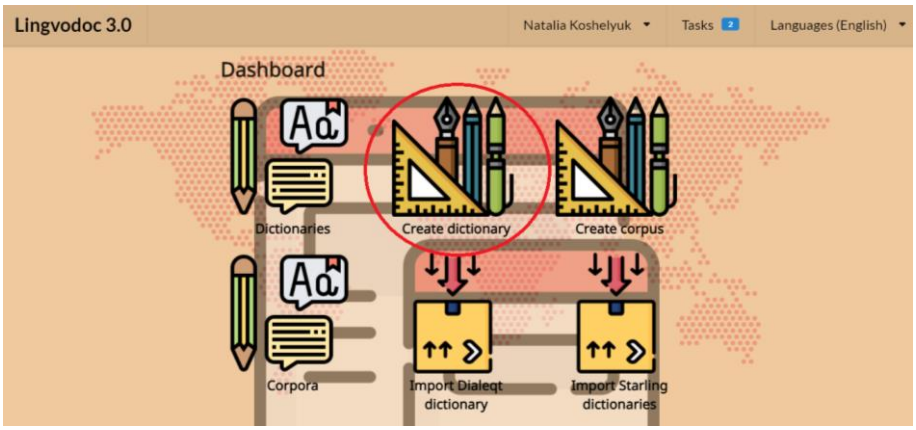
There are several ways to create a dictionary:

1) converting a dictionary in Excel format through the corpus of the required language or the dialect already created in LingvoDoc – in this case, you can find the corpus of the language of interest with the help of the search engine and create a dictionary by using the option “Convert to dictionary” (Pic. 1);



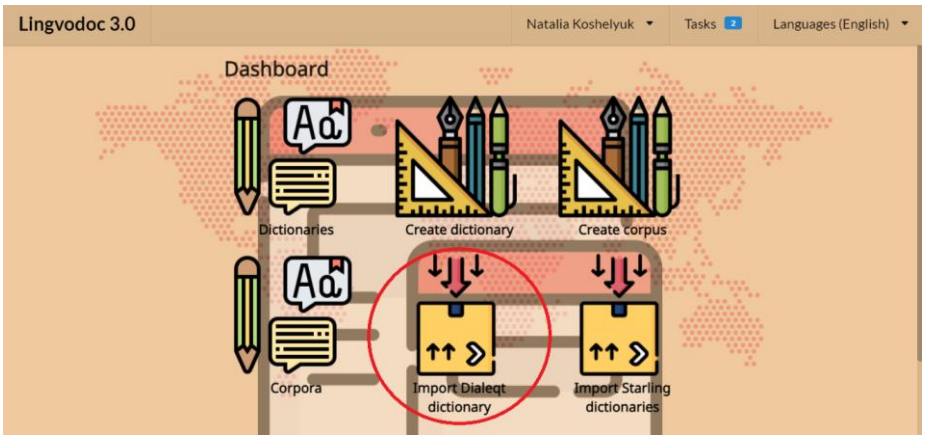
Pic. 1. Converting a dictionary from the corpus.

2) online – using the option “Create dictionary” (Pic. 2);



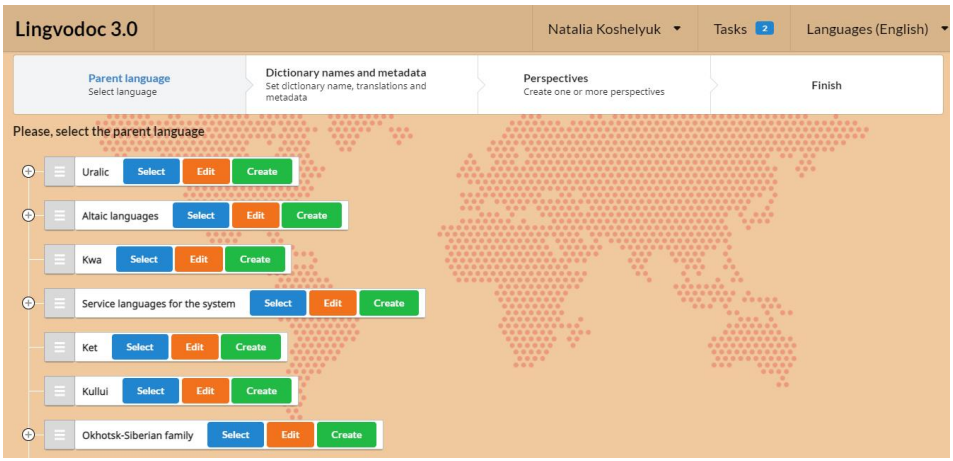
Pic. 2. Creating a dictionary directly on the platform.

3) adding a dictionary created in Word using the option "Import Dialect Dictionary" (Pic. 3);

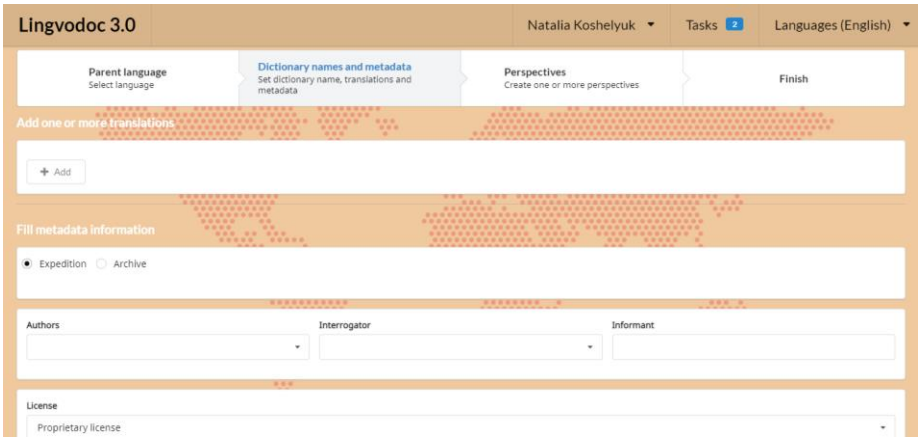


Pic. 3. Creating a dictionary using Word.

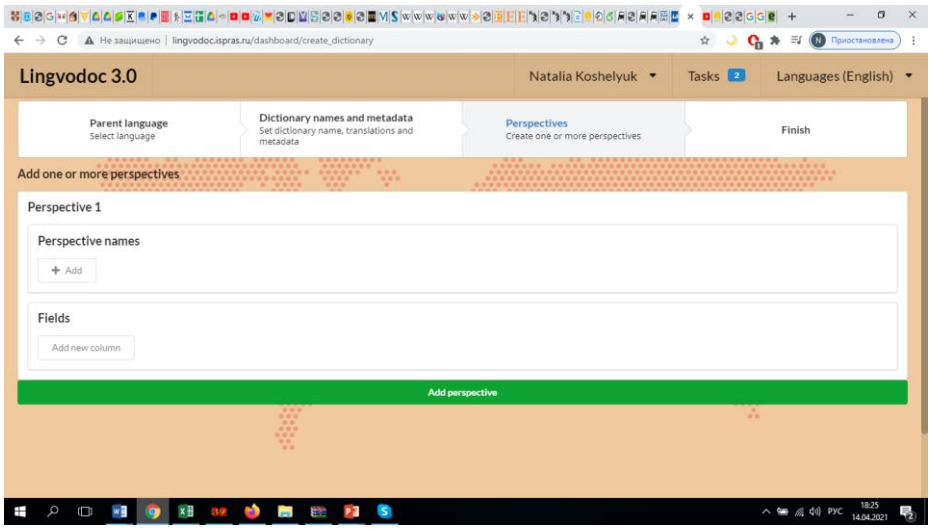
If you create a dictionary online on LingvoDoc website or add a Word file, you need to select a language family, a language and a dialect (Pic. 4), fill in the metadata – archival or field data, the year the source was created or the year the data was collected, the author's name, etc. (Pic. 5), and add perspectives (Pic. 6). Once all these steps are completed, the dictionary will be displayed in the database.



Pic. 4. Selecting the desired language or dialect.



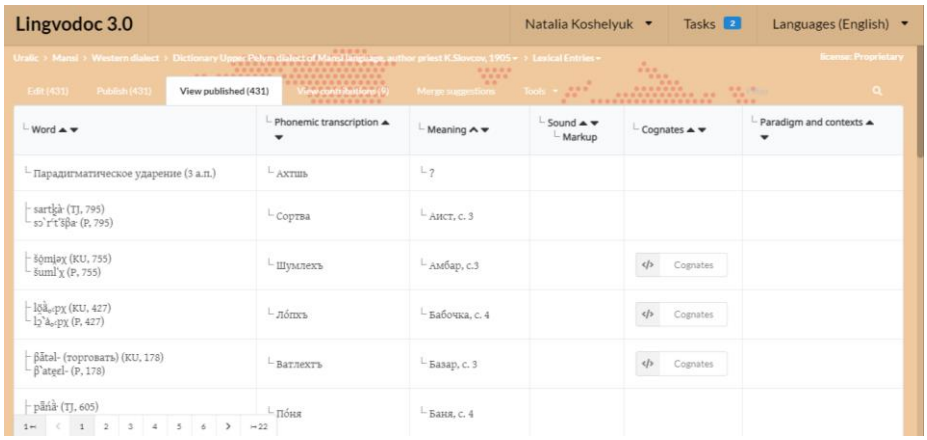
Pic. 5. Filling in the metadata.



Pic. 6. Adding perspectives.

You can find the dictionary you've created by typing in its name in the search engine on the main page or in your personal account under "My files". The user can add restrictions or extensions to it, if necessary, for example: hide the source, allow access to all registered users or allow access to specific persons, and so on.

For further research, open your dictionary and fill in the required fields: "Phonemic transcription", "Meaning", etc. (Pic. 7). In the field "Phonemic transcription", enter the word from the archive; in the field "Meaning", indicate the Russian translation of the word and its localization, i.e. the page.



Pic. 7. Filling in the required fields of the dictionary.

After filling in the required fields, proceed to adding parallels from other reliable sources which contain the information on the language or dialect under study (in our case, the data of A. Kannisto's dictionary correspond to the Western Mansi dialect). It has to be done for the purposes of: 1) checking the accuracy of the new dictionary's data and 2) checking whether there are any innovations, archaisms, discrepancies in the data of both researchers and other interesting phenomena. After analyzing the graphs of the sources under study (Pic. 8), as a result of comparison, the researcher sees a number of unambiguous (Pic. 9) and ambiguous phonetic correspondences and draws certain conclusions about the nature of their appearance and its significance for a specific research and study of the disappearing Mansi language in general. At the same time, the study of ambiguous phonetic correspondences has a number of peculiarities: to identify the reasons for their appearance, it is necessary to make a complete list of varying words, add correspondences from other Mansi dialects and connect them to external etymologies from other languages and dialects. To implement these steps, the users of the LingvoDoc platform can use the "Cognates" option (Pic. 10).

How to add etymological connections to words in the dictionary under study:

1) Enable the editing mode of the dictionary and go to the "Cognates" field;

2) Next, you will switch from the "Cognates" field to "Add connection" (Pic. 11). After filling in the appropriate word in the open field, LingvoDoc will automatically search for etymologies in the database. The results must be reviewed and checked for possible errors and further linked by cognates by means of selecting the required elements (Pic. 12). It is also important to understand that the word you are looking for might not have etymological connections with other dialects and languages.

[Словцов 1783]	[Kannisto 2013]
в	*β*
г	*γ*
й	*j̱*
к	*k*
л	*л, l*
м	*m*
н	*n*
нг	*ŋk*
п	*p*
р	*r*
т, д	*t*
х	*χ*
с	*s, ś*
ш	*š, ś*
ч	*ś*

Pic. 8. A comparison of consonant graphemes in the dictionaries of Father K. Slovtsov and A. Kannisto.

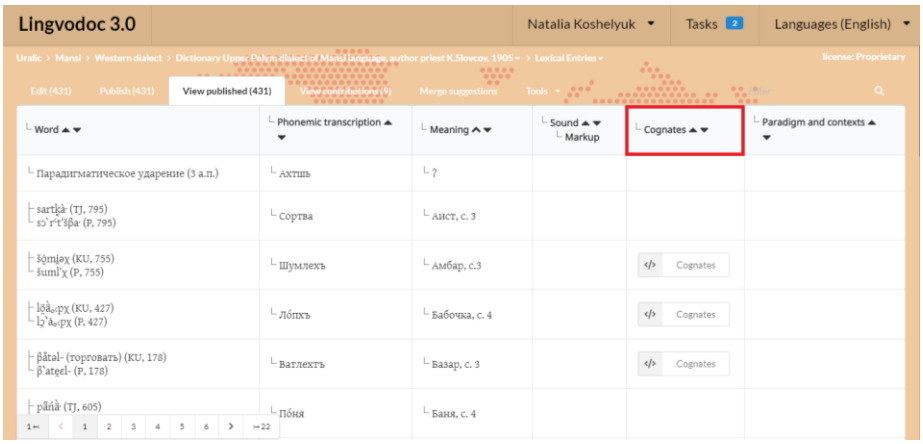
пелым. *в*	**пелым. *β***
[Словцов 1905]	**[Kannisto 2013]**

пелым. *Ватлехтъ* ‘базар’ [Словцов 1905: 3], *βàtɐ̀èl-* ‘базар’ [Kannisto 2013: 178];
пелым. *Вэта* ‘берег’ [Словцов 1905: 3], *βɐ̄tà* ‘берег’ [Kannisto 2013: 177];
пелым. *Вышка* ‘бык’ [Словцов 1905: 4], *βɐ̌škà* ‘бык’ [Kannisto 2013: 110];
пелым. *Вондартъ* ‘выдра’ [Словцов 1905: 5], *βɔntᵊrt* ‘выдра’ [Kannisto 2013: 154];
пелым. *Вýрнъ* ‘двор’ [Словцов 1905: 8], *βurn* ‘двор’ [Kannisto 2013: 96];

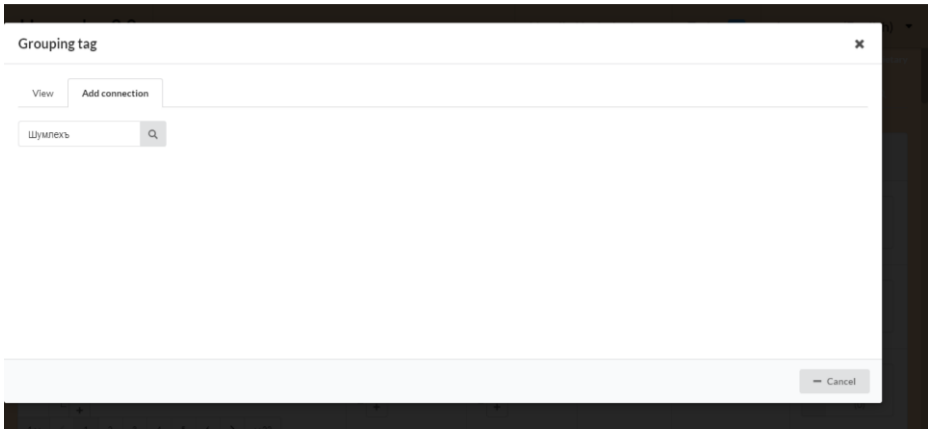
пелым. *г*	**пелым. *γ***
[Словцов 1905]	**[Kannisto 2013]**

пелым. *Ригъ* ‘жарко' [Словцов 1905: 11], *riɣ* ‘жарко’ [Kannisto 2013: 979];
пелым. *Егъ* ‘отец’ [Словцов 1905: 19], *iɛɣ* ‘отец’ [Kannisto 2013: 188];
пелым. *Игъ* ‘ночь’ [Словцов 1905: 18], *iiɣ* ‘ночь’ [Kannisto 2013: 119];
пелым. *Нэгъ* ‘ты’ [Словцов1905: 27], *nɛ̄ɣ nɛɣ* ‘ты’ [Kannisto 2013: 493];

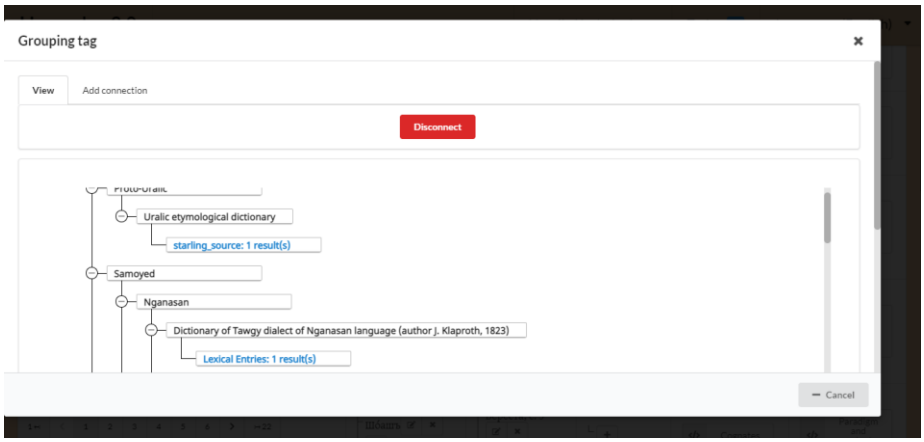
Pic. 9. A number of unambiguous correspondences.



Pic. 10. The option "Cognates".



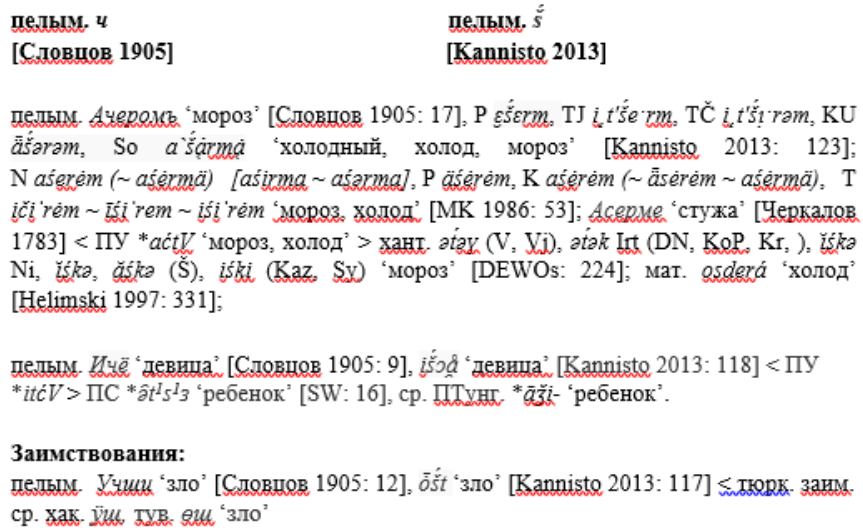
Pic. 11. The option "Cognates", add connection.



Pic. 12. A word with related etymologies.

It is worth noting that this kind of work took linguists quite a lot of time in the past: searching for etymologies in the dictionaries and other paper (or electronic) media could take weeks and even months.

By connecting all the words of the dictionary under study by etymological links, we can check the detected discrepancies in the data of the graphic and phonetic analysis we are conducting. As a result, the researcher will have lists with added etymological parallels (Pic. 13) that will help to conclude what a certain discrepancy is linked to.



пелым. ч **пелым. ś**
[Словцов 1905] **[Kannisto 2013]**

пелым. *Ачеромъ* 'мороз' [Словцов 1905: 17], P *ɛśśerm*, TJ *i̯ t'śe·rm*, TČ *i̯ t'śi̯·rəm*, KU *ãśərəm*, So *a`śą̊rmą̊* 'холодный, холод, мороз' [Kannisto 2013: 123]; N *aśęrėm (~ aśėrmä) [aśirma ~ aśərma]*, P *äśėrėm*, K *aśėrėm (~ ãsėrėm ~ aśėrmä)*, T *iči̯'rėm ~ ĩśi̯'rem ~ iśi̯'rėm* 'мороз, холод' [MK 1986: 53]; *Асерме* 'стужа' [Черкалов 1783] < ПУ **aćtV* 'мороз, холод' > хант. *ə̑təɣ* (V, Vj), *ə̑tək* Irt (DN, KoP, Kr,), *ĭśkə* Ni, *ĭśkə*, *äśkə* (Š), *iśki* (Kaz, Sy) 'мороз' [DEWOs: 224]; мат. *osderá* 'холод' [Helimski 1997: 331];

пелым. *Ичё* 'девица' [Словцов 1905: 9], *i̯śɔ̑ą̊* 'девица' [Kannisto 2013: 118] < ПУ **itćV* > ПС **ə̑t^{1}s^{1}ə* 'ребенок' [SW: 16], ср. ПТунг. **ä̆ži-* 'ребенок'.

Заимствования:
пелым. *Учшц* 'зло' [Словцов 1905: 12], *ə̑śt* 'зло' [Kannisto 2013: 117] < тюрк. заим. ср. хак. *ӱш*, тув. *өш* 'зло'

Pic. 13. Series of unclear correspondences with added etymological connections.

CONCLUSION

As can be seen from the above, with the help of the automated data analysis algorithms built into LingvoDoc, the modern linguists can analyze sound data at a completely new level that provides a more detailed and prompt results of data processing with an ability to check the obtained results as opposed to the methods and technology of the experimental linguistic research of the previous years.

REFERENCES

1. LingvoDoc Homepage, http://lingvodoc.ispras.ru/. Last accessed 14.04.2021.