EasyChair Preprint

№ 6242

LingvoDoc: Working with Text Corpora

Natalia Koshelyuk

August 6, 2021

LingvoDoc: working with text corpora[1]

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

Abstract. The paper is a continuation of a series of articles about work and user experience on the LingvoDoc linguistic platform. Step-by-step consideration is being given to ways of creating text corpora on a platform, as well as to the representation of new opportunities for carrying out linguistic research at a new and improved level.

Keywords. LingvoDoc, Corpus, endangered languages, data mining, linguistics

1 INTRODUCTION

A text corpus (corpus linguistics, language corpus) is a set of electronic data intended for solving specific linguistic problems, collected according to certain principles, marked up in accordance with a certain standard and provided by a special search engine. Also, any collection of texts combined by some common feature (language, genre, author, period of creation) can be considered as a corpus. The usefulness of creating text corpora is explained by 1) representation of linguistic data in a real context; 2) a rather large representation of data (in case of a large corpus); 3) the possibility of repeated use of the once created corpus for different linguistic tasks, etc. Corpora linguistics works closely

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

with computer linguistics in the points of development, creation and use of text corpora. Corpora can be used to obtain a variety of information and statistics on linguistic and linguistic units of a language. In particular, they can provide data on the frequency of word forms, lexemes, grammatical categories, changes in context at different times, etc. A representative set of linguistic data for a certain period makes it possible to study the dynamics of the processes of the language lexical composition change, to conduct analysis of lexical-grammatical characteristics in different genres and in different authors. Corpora are also intended to be a source and a tool for producing a variety of historical and modern dictionaries, and can be used to construct and refine grammar and to teach language.

The first linguistic text corpora appeared in the 1960s, and by the beginning of the 21st century such databases have been created for many languages of the world. Corpora on Altai and Uralic languages and dialects, including those that have already disappeared or are on the verge of extinction are widely provided on LingvoDoc (for more information see [Kosheliuk 2021]).

This article provides a consistent description of steps to create and work with text corpora in LingvoDoc.

2 HOW TO MAKE COPRUS

Work with sources on the LingvoDoc platform always begins with registration in the system or login to your account. After this step, the user will be able to work with the full linguistic base of the platform (in some cases with the prior permission of the author of a dictionary or

corpus) and use all of its options, including the creation of its own corpus.

The basic of the future corpus and the material on which further research is based are often the archived data discovered by researchers or field data collected during the expeditions. In our case, an example of the use of the LingvoDoc platform was the «Gospel of Matthew 1868» of the disappearing Mansi language - the Aboriginal language of Siberia, presented on the website LingvoDoc.[2]

Work with the corpus can be started in two ways:

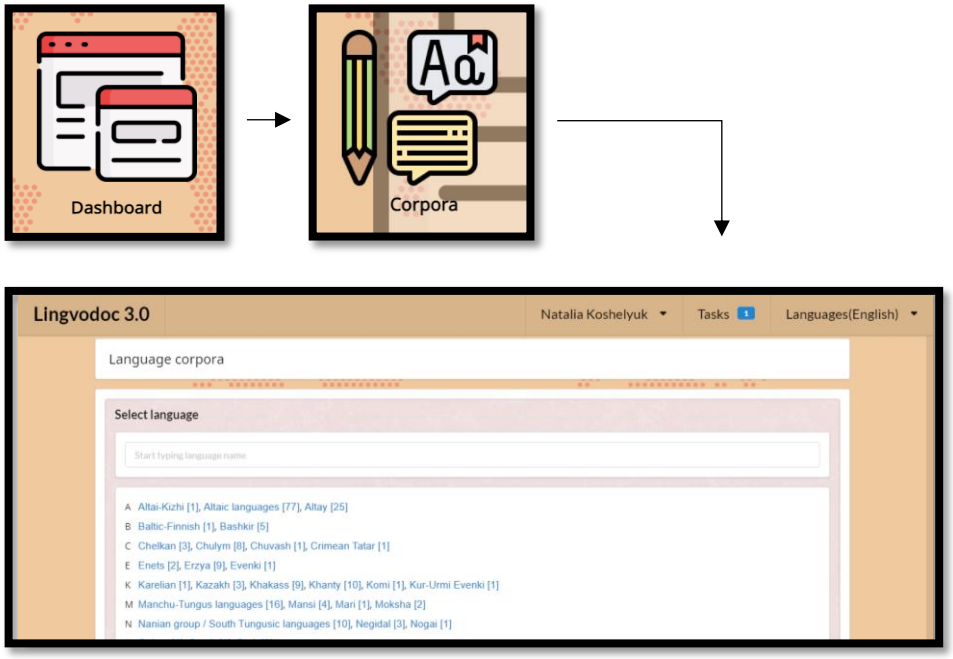
1) The opening of the existing one by authorization to the site, by entering the Language Database > Language Corpora;

2) Creating a new corpus via Dashboard > Create Corpus > Corpora[3] (Picture 1).

Picture 1. A sequence of steps to create a corpus

Once you have created your own corpus, you can start working directly with texts. To do this, you must open your corpus or search for an existing one on LingvoDoc. The user-created corpora are stored in his personnel office, to which the transition is made as follows:

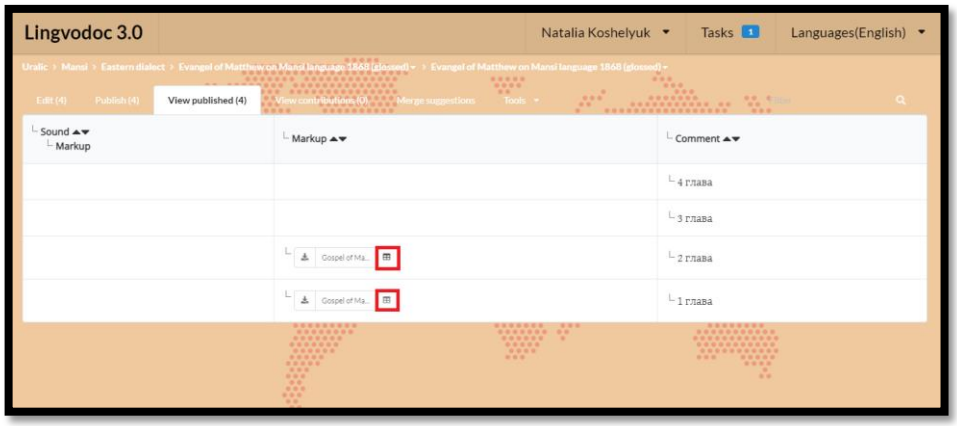
[2] See link: http://lingvodoc.ispras.ru/dictionary/1146/10/perspective/1146/13/view.
[3] Corpora displays user-created corpuses.



3 ALGORITHM FOR FURTHER ACTION

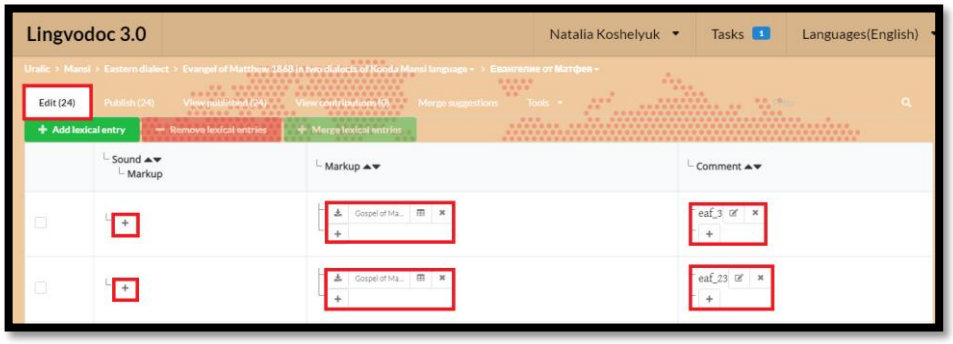
After opening the desired corpus, the user can explore the language material at any convenient time, without a session time limit and independently of other users (Picture 2).

Picture 2. Opening the corpus in the system



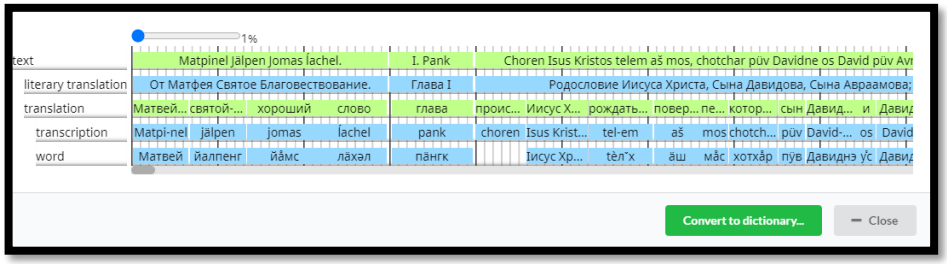
When working with corpora, linguists can also check added text, edit it, attach sound files and perform other required actions (Picture 3). These actions can be performed both in the corpus and in the corpus of another user - if there are added rights requested personally from the editor-in-chief of the platform and/or author of the required corpus.

Picture 3. Corpus edit field



After performing all necessary corpus manipulations, it will look like this (Picture 4):

Picture 4. Finished corpus at LingvoDoc

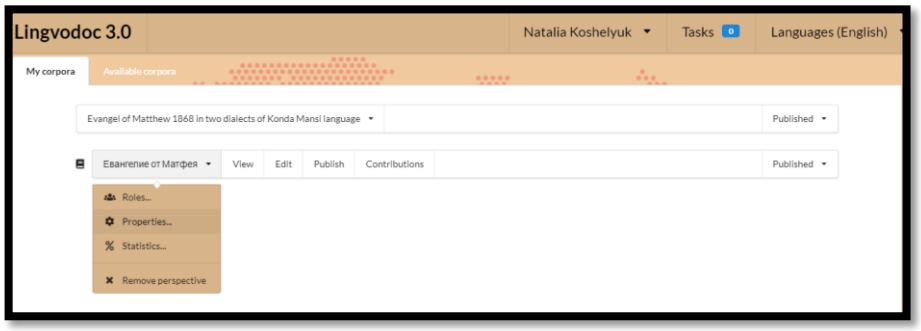


It is not enough to have an array of texts to solve different linguistic problems. Texts are also required to contain, in a variety of ways, additional linguistic and extra-linguistic information. This is how the idea of a demarcated corpus arose in corpus linguistics. **The markup**

(tagging, annotation) consists of attributing special marks (tag, tags): external, extralinguistic (information about the author and information about the text: author, title, year and place of publication, genre, subject; information about the author may include not only his name, but also age, sex, years of life and much more. This encoding of information is called meta-markup), structural (chapter, paragraph, sentence, word form) and linguistic proper, describing lexical, grammatical and other characteristics of text elements. The set of this metadata largely determines the capabilities offered by the corpora to the researchers. The selection of these data should be guided by the objectives of the study and the needs of the linguists, as well as the possibility of adding non-core topics to the text.

Extra-linguistic information on LingvoDoc is not displayed directly when opening the corpus, but it can be added in the personnel office in the special Properties (Picture 5):

Picture 5. The field for adding extralinguistic information to the corpus

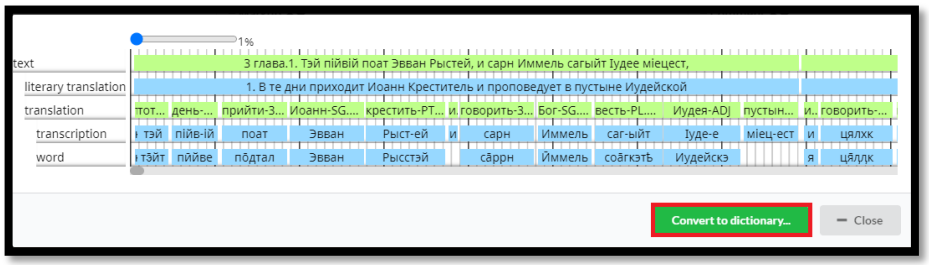


The following fields are usually displayed directly in the corpus markup (Picture 4) on LingvoDoc:

- text – phrase from the monument;
- literary translation – Russian translation;

- translation – recording a gloss;
- transcription – indicators;
- word – a parallel from a related language

Concordance can also be formed from any corpus:[4]

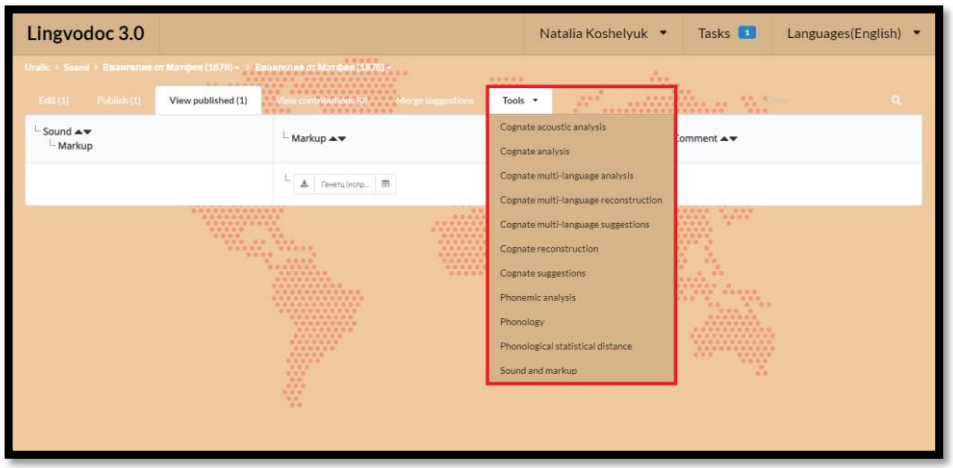


NEW DICTIONARY

Moreover, the options available when working with text corpora are similar to those for working with dictionaries (Picture 6): cognate acoustic analysis, cognate analysis, cognate multi-language analysis, cognate reconstruction, cognate suggestions, phonemic analysis, phonology etc.

[4] Concordance is a list of all uses of a given word in context with links to the source.

Picture 6. List of options available when working with corpus on LingvoDoc

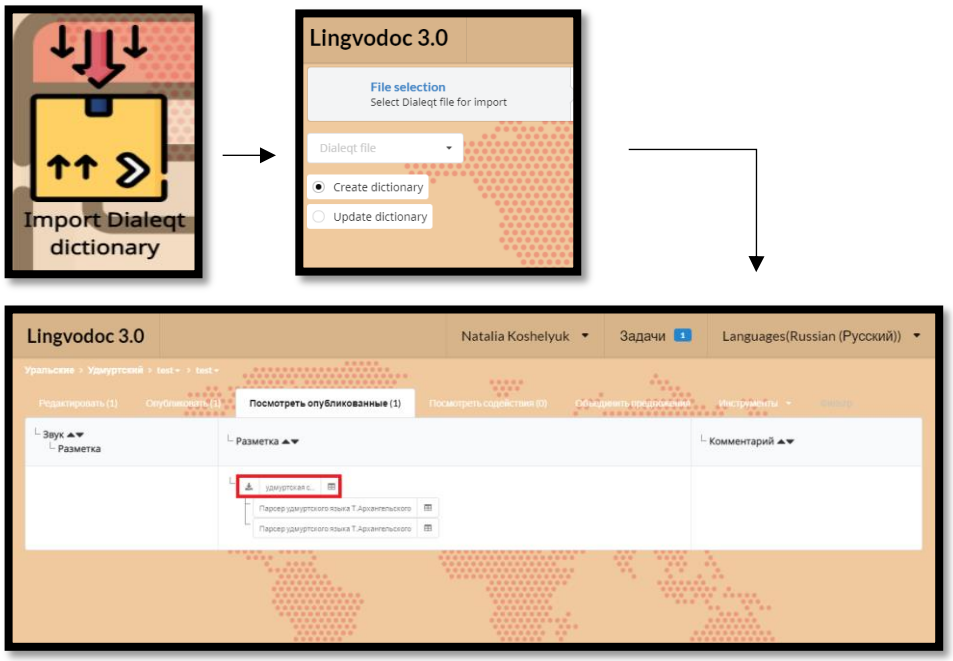


IV NEW OPPORTUNITIES FOR WORKING WITH CORPORA

Until recently, working with the LingvoDoc corpora was a bit of laborious: the corpora was originally formed by the Elan program[5], and a mechanism for building corpora directly in LingvoDoc was developed (method described above) in order for a linguist to start researching texts. However, since 2020, a new option has been launched using Timothy Arkhangelsky (Hamburg) code, which is to load text on LingvoDoc without having to create markup - creating parsers[6]. This option features the ability to type any text in Word and load it in *.odt* format (Picture 7).

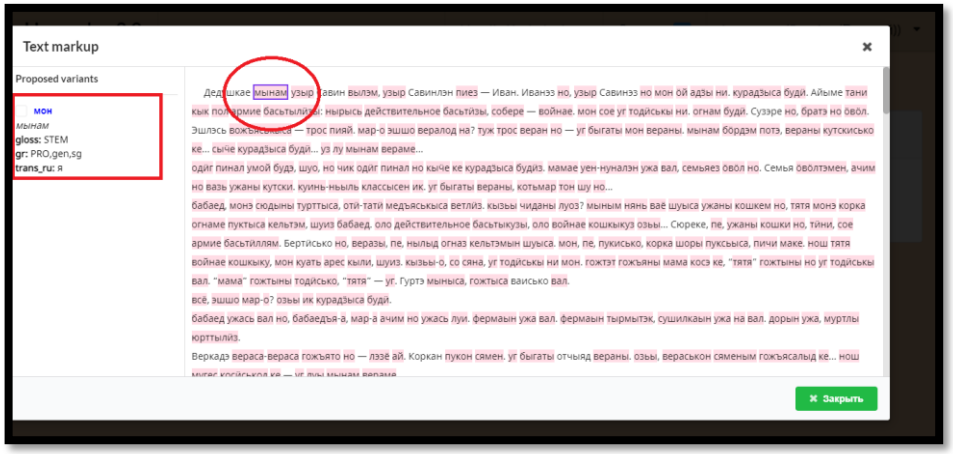
[5] Elan is a computer software, a professional tool for manual and semi-automatic annotation and decryption of audio or video recordings.
[6] A parser is software that extracts certain pieces of information from a data set.

Рисунок 7. Creating a parser on LingvoDoc



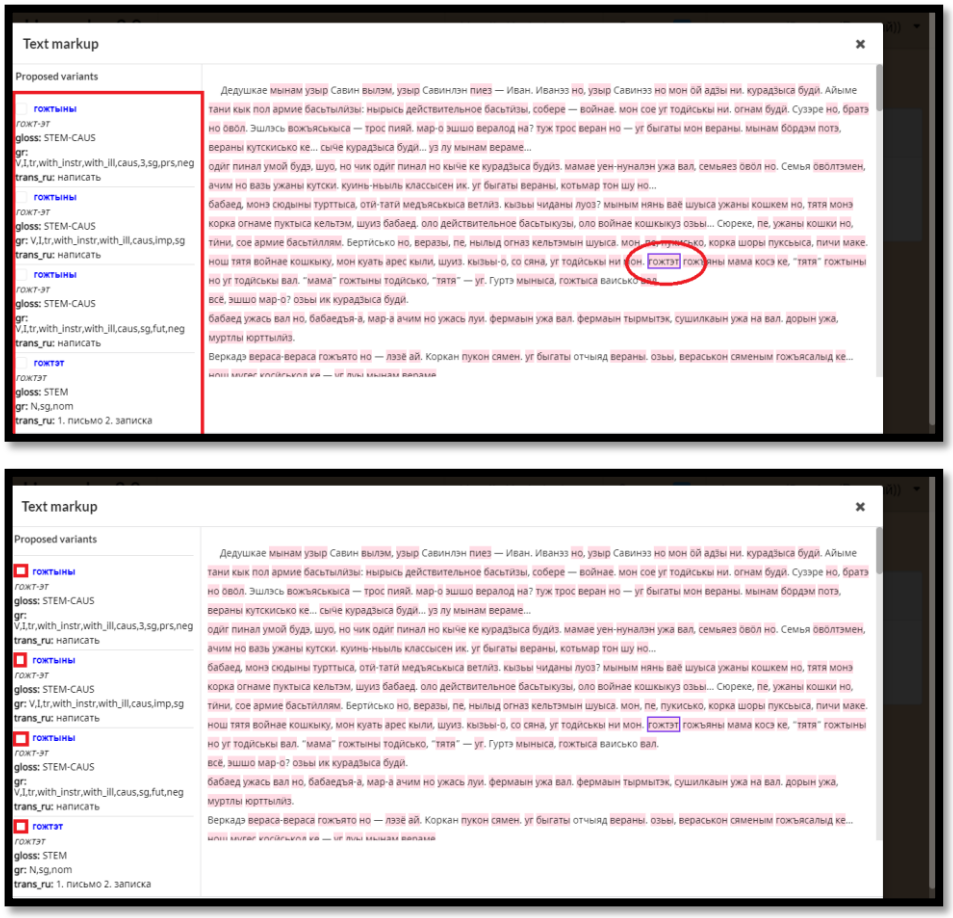
In the text processed with the parser, the markup occurs automatically (Picture 8):

Picture 8. Implementation of automatic parser markup



In case of ambiguous markup, you can use hovering the cursor on a glossed word to remove the homonymy (similarity of words by sound or spelling, but different meaning) in the text (Picture 9):

Picture 9. An example of removing homonymy in a parser



CONCLUSION

The features for working with text corpora provided by Lingvo-Doc platform allow not only to create databases in a more convenient way, but also to use on this material all the options available when working with dictionaries.

In addition, the system's built-in option of adding parsers, due to the possibility of a Word text set, automatic markup and removal of homonymy, provides a more detailed, fast, but at the same time reliable scientific result.

REFERENCES

1. LingvoDoc. Homepage, http://lingvodoc.ispras.ru/. Last accessed 21.07.2021
2. The National Corpus of Russian languages. Homepage, https://ruscorpora.ru/new/corpora-intro.html/. Last accessed 19.07.2021