

LingvoDoc: Working with Text Corpora

Natalia Koshelyuk

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

August 6, 2021

LingvoDoc: working with text corpora¹

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

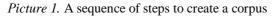
Work with sources on the LingvoDoc platform always begins with registration in the system or login to your account. After this step, the user will be able to work with the full linguistic base of the platform (in some cases with the prior permission of the author of a dictionary or corpus) and use all of its options, including the creation of its own corpus.

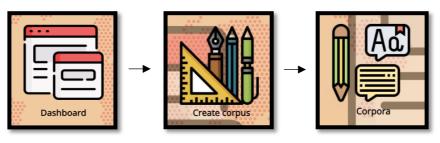
The basic of the future corpus and the material on which further research is based are often the archived data discovered by researchers or field data collected during the expeditions. In our case, an example of the use of the LingvoDoc platform was the «Gospel of Matthew 1868» of the disappearing Mansi 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 > Corpora³(Picture 1).

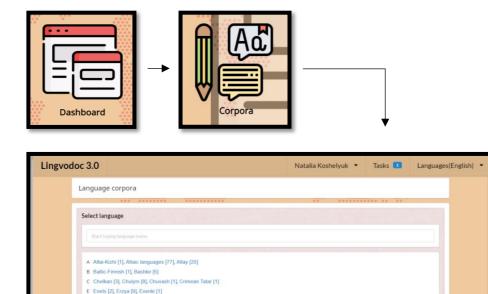




Once you have created your own corpus, you can start working directly with texts. To do this, you must open your corpus or search for an existing one on LingvoDoc. The user-created corpora are stored in his personnel office, to which the transition is made as follows:

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

K Karelian [1], Kazakh [3], Khakass [9], Khanty [10], Komi [1], Kur-Urmi Evenki [1] M Manchu-Tungus languages [16], Mansi [4], Mari [1], Moksha [2] N Nanian group / South Tungusic languages [10], Negidal [3], Nogai [1]

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).

Lingvodoc 3.0	Natalia Koshe	elyuk • Tasks 1 Languages(English) •
Uralic > Mansi > Eastern dialect > Evangel of Mat	thew on Manul Language 1868 (closeed) - > Evangel of Matthew on Manul Language 1868	1993
Edit (4) Publish (4) View published (4)) View contributions (0) Merge suggestions Tools +	
└─ Sound ▲▼ └─ Markup	L Markup 🔺	└ Comment ▲▼
		^L 4 глава
		^L 3 глава
	Gospel of Ma.	^L 2 глава
	Cospel of Ma_	∟ _{1 глава}

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
------------	--------	------------

Lingvoo	doc 3.0		Natalia Koshelyuk 💌	Tasks 💶	Languages(English)
Uralic > Marca Edit (24) + Add lexis	Publish (24) View published (24)				٩
	└─ Sound ▲▼ └─ Markup	L Markup 🔫	L	Comment 🗪	
	•	↓ Gospel of Ma ⊞ ×		eaf_3 ≥ ×	
	+	A Cospel of Ma_ III X		eaf_23 ⊠ × +	

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

Picture 4. Finished corpus at LingvoDoc

раамова;
и Дави
os David
ус Дави,
4

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

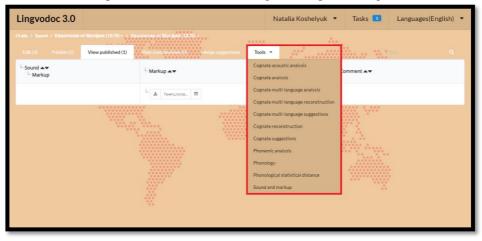
Concordance can also be formed from any corpus:⁴

	1%	
text	3 глава.1. Тэй пійвій поат Эвван Рыстей, и сарн Иммель сагыйт Іудее міецест,	
literary translation		
translation	тот день прийти-3 Иоанн-SG крестить-РТ и. говорить-3 Бог-SG весть-PL Иудея-ADJ пустын и. говори	
transcription	і тэй пійв-ій поат. Эвван. Рыст-ей и сарн. Иммель саг-ыйт. Іуде-е міец-ест и цял	іхк
word	ітэйт пийве подтал Эвван Рысстэй саррн Йммель соагкэть Иудейскэ я цял	
	Convert to dictionary	Close
	•	
	NEW DICTION A DV	

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

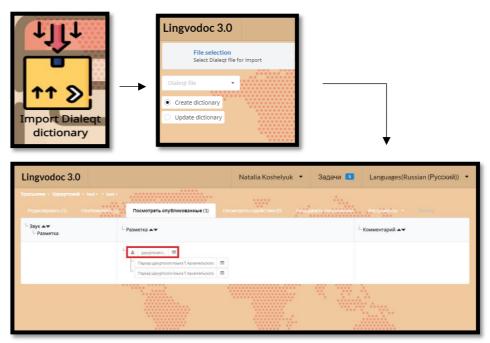
IV NEW OPPORTUNITIES FOR WORKING WITH CORPORA

Until recently, working with the LingvoDoc corpora was a bit of laborious: the corpora was originally formed by the Elan program⁵, 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 parsers⁶. This option features the ability to type any text in Word and load it in *.odt* format (Picture 7).

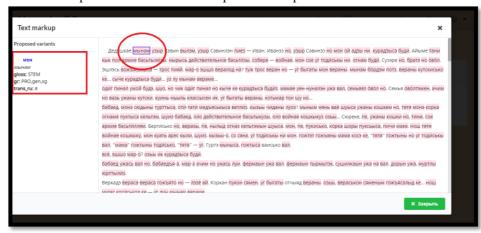
⁵ Elan is a computer software, a professional tool for manual and semi-automatic annotation and decryption of audio or video recordings.

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

Text markup	×	
Proposed variants Prostainta rowship rowship rowship rans, rut, annicata rowship rowsh	Дедушкае мынам узыр Севин вылам, узыр Савинлан пиез — Иван. Иваназ но, узыр Савиназ но мон ой адзы ни, курадзыса будй. Айыме тани кык пол армие басытылизы: нырысы действительное басытазы, собере — войнае, мон сое уг тодиськы ни, огнам будй. Сузэре но, братз но овол. Эшилсь вожаясысас — трос пияк, мар-о эшшо вералод на? тух трос веран но — уг быгаты мон вераны. мынам бордам потз. вераны кутскисько ке сыче курадзыса будй уа лу мынам вералод. одит пиная умоб буды. шус но ики одит пинал монеке ке украдиса будиз. мамае уен нуналан ужа вал, семьяез овол но. Семья оволтамен, ачим но вазь ужаны кутски, куинь-нышть классысен их. уг быгаты вераны, котьмар тон шу но бабаед, коно содыны тутски, куинь-нышть классысен их. уг быгаты вераны, котьмар тон шу но бабаед, коно содыны тутски, пит-тати медъяськыга ектата, кызыы чиданы луоз? иыным наны вае шумса ужыны кошкем но, тятя моны корка огнаме пуктыс кектилы. но иска бабаед, опо действительное басытызузы, опо войнае кошкыхую саы Сокры, пе укины кылы но, тими, кое армие баскитилия. Сектокьон кана, кызы бы кользы кираны и бы болгок, кураны кошкем но, татя моны нош тятя войнае кошкыку, мон куать арес кыли, шуиз, кызыв-о, со сина, уг торудны кошкы ку, кира кала, сок кулама, корка, корка шоры пуктыска, пини маке нош тятя войнае кошкыку, мон куать арес кыли, шуиз.	
rowrbitbi owr-ar joss: STEM-CAUS jr: ,[tr,with_instr,with_ill,caus,sg,fut,neg rams_ru; Hanucatb	но ут тодатськи вал. чамаат гожпыны тодатсько, "тятя" — уг. Гуртэ мыныса, гожтыса вамсько ма все, вшшо мар-07 овы их курадныга буди. бабаед ужась вал но, бабаедъя-а, мар-а ачим но ужась луи, фермаын ужа вал, фермаын тырмытэк, сушилкаын ужа на вал. дорын ужа, муртлы коттыких.	
rowtat oxrar loss: STEM r: N.sg.nom	Верхад» вераса вераса гожъято на — иззе ай. Коркан пукон сямен. ут быгаты отчыяд вераны, озвы, вераськон сяменым гожъясалыд ке… Шли илие инигисти и — от якы изыкан вялация	
rans_ru: 1. письмо 2. записка		
rans_ru: 1. письмо 2. записка Text markup	x	M)) -
		й)) -
Text markup	Дедушкае мынам узыр Савин вылзм, узыр Савинлэн пиез — Иван. Иванз но, узыр Савинз но мон бй адзы ни. курадзыса будй. Айыме Танк кык пол армие бастылахы: нирисс действительное басыткы, собере — войнае. кон сое ут тодиськы ни. огнам буди. Сузое но, братз но бвол. Эшлэсь вожнаскысса — трос пияй. мар о эшшо вералод на? туж трос веран но — ут быгаты мон вераны. мынам бордэм потэ, вераны кутскиских ке сыче курадзыса буди у ли ушнам вераме одит пиная умой буда, шуо, но чик одит пинал но кыче ке курадзыса будиз. мажае уен нуналэн ужа вал, семыяез бвол но. Семья оволтэмен,	· ((۲
Text markup roposed variants roposed variants rorstwike ses: STEM-CAUS ritzwith_instr.with_ill_caus,3,sg,prs.neg ans_rut_namcats rows-array ses: STEM-CAUS v.U.tz.with_instr.with_ill_caus,imp.sg	Дедушкае мынам узыр Савин вылзм, узыр Савинлэн пиез — Иван. Иванээ но, узыр Савинээ но мон ой адзы ни, курадзыса будй. Айыме Танк кык пол армие басьтылазы: нырысь действительное басытазы, собере — войнае, кон сое ут тодиськы ни, олнам буди. Сузэре но, братэ но бвол. Эшлись вожавськаса — трос пияа, мар э зшис у по изынам вералод на? туж трос веран но — ут быгаты мон вераны. нынам борды сузэре но, братэ но бвол. Эшлись вожавськаса — трос пияа, мар э зшис у по изынам вералод. на? туж трос веран но — ут быгаты мон вераны, нынам борды потэ, вераны кулсковк ке сыне курадзыса буди. Э ли изынам вераны. одит пинал умой будь шо, не чик одит пинал но кыче ке курадзыса будиз. мамае уен-нуналэн ужа вал, семьяез бвол но. Семья оволтэмен, ачим но вазь ужины кулсок, кумы-тызлы какссыссн ик. ут быгаты вераны, котымар тон шу но бабаед, монэ содыны турттыса, отл-тати медывськаса веглал, кызыы чиданы укоэ? мыным наны вае шуыса ужаны кошкин но, татя монэ корка олжает мулсыс какстык кума бабев, од по деятствителен бе больтыке кошкака должа колькака тоны шуюз? тины, сое армие басытилям. Бергизсько но, веразы, пе, нылыд огназ кельтэмын шуыса. мон, пе, пукисько, корка шоры пуксыса, тини маке.	(M))
Text markup roposed variants Trowniki Sari Jar Soss: STEM-CAUS Tr. Lirwith, Instr.with, III, caus, 3, sg, prs. neg	Дедушкае мынам узыр Савин вылзм, узыр Савинлэн пиез — Иван. Иванз но, узыр Савинзэ но мон бй адзы ни. курадзыса будй. Айыме Танк как пол армие бастыллах: нирись действительное басытлах, собере — войнае. кон сее у тодиськы ни, огнам буди. Сузоре но, бортз но бвой, зилась вокласкажса — трос пияй. мар о зшиш вералод на? тух трос веран но — ут бигаты мон вераны. Кинам бордом потэ, вераны кутскискок как. сыче курадзыса буди. уз лу изнам веранод. на? тух трос веран но — ут бигаты мон вераны. мынам бордом потэ, вераны кутскискок как. сыче курадзыса буди. уз лу изнам веране. одаг пиная умой будь шуо, но чик одаг пинал но кыче ке курадзыса будиз. камае уен-нуналэн ужа вал, семляез беол но. Семля оволтэмен, ачим но вазь ужаны кутски, кумин-пьыль каксоскен их. ут бигаты вераны. котыхар тон шу но бабеае, моно содины турттыса, кота тата медалескае ветлять, каконы чираны вицимае вшумса уханы кошкем но, татя монн корка огнаме тухтыса кельтэм. шуиз бабаед, оло действительное басытынузы, оло войнае кошкыкуз озы Сореке, пе, ужаны коми но,	())

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