O.D.Borisenko, Ju.V.Normanskaja

Dictionaries on Samoyedic languages and LingvoDoc software system for collaborative work on dictionaries and online publishing[1]

[1] This paper was financially supported by Ministry of Education and Science of the Russian Federation (grant No. 14.Y26.31.0014.).

Abstract

The LingvoDoc software system (http://lingvodoc.ispras.ru/) is a recent piece of computer technologycreated especially for describing and recording endangered languages. This paper illustrates how the authors used LingvoDocto bring the description and analysis of Samoyedic languages to anew level with easily verifiable results. To date, the project has produced:

1. **10 online audio-dictionaries** covering all presently living dialects of Samoyedic languages;
2. **6 audio-dictionaries and concordances** based on archival data, with etymological connections traced to entries in the audio-dictionaries.

The hosting of these materials through LingvoDoc gives scholars the opportunityto analyze the information in a convenient format — both by etymological groupand separately for each dialect at a given period in the history, with the option to trace changes within each dialect.

1 Field materials

In 2011-2015 we processed our collected field data and completed the creation of dialectal audio-dictionaries for the following Samoyedic languages: 1) Nenets (dictionaries on four dialects: Tundra: Yamal, Kanin, Gydan, Bol'shezemel'skij), 2) Enets (dictionaries on two dialects – Enets and Forest), 3) Nganasan (dictionaries on two sub-dialects: Ustj-Avam and Volochanka), 4) South-Selkup (Narym, Ket dialect). In practical terms,all of these dialects are endangered. There are no more than ten native speakers of Enets and Nganasan and only one fluent speaker of the Narym dialect of Selkup. Speakers of all dialects eagerly anticipate the creation of audio-dictionaries, available for listening and correction (after approval by the site administrators) online. It is likely that we are currently facing the last opportunity to work with speakers of these endangered languages in such a mode, where each speaker has internet access, can look through the dictionary, and can send comments and corrections.

2 Archival materials

At present, there exist numerous (in some cases, very detailed) descriptions of contemporary Samoyedic languages. There is also a generally accepted reconstruction of the proto-language [Janhunen 1977, Mikola 2004, Helimski 2000]. However, all of the Samoyedic languages are, at best, only recently literate, and many do not exist in written form at all; therefore, information about the first recordings of these languages has been poorly studied up to present but is very important for further research.

1. **In the 19th century,**the first books in the Nenets and Selkup languages were created by the Russian Bible Society’s Translation Committee. The first writing systems developed for these translations were based on the Cyrillic alphabet and aimed to adequately reflect the sounds of the spoken language. Now, these books are available on-line (http://uralica.kansalliskirjasto.fi/). It is difficult to overestimate the importance of these

first large textual and grammatical sources for the history of the Samoyedic languages. Our preliminary analysis found thatthe Selkup dialectal features reflected in the Bible Society's texts fully coincide with M. A. Castrén's (18thcentury) and A. P. Dulzon's (20thcentury) materials. This concurrence among archival sources indicates that no major changes have taken place in the Selkup dialects in the last 200 years; thus, all sources on the dialects are valuable and should be fully included into scientific circulation. The first Nenets-Russian dictionary with etymological connections is nowavailable online on lingvodoc.ispras.ru.

2. **In the 1970s,**scientists from Russia (especially Novosibirsk) collected numerous audio-records of Samoyedic languages. Some of these recordings —3 from Nganasan (Ustj-Avam and Vadeeevo dialects), 1 from Enets and 1 from Selkup (Middle-Ob' dialect) are also available online on lingvodoc.ispras.ru.

3 Research methods

As mentioned before, this project makes use of the unique software system LingvoDoc (lingvodoc.ispras.ru; for an overview of this system, see below).LingvoDoc allows for the creation of online multimedia dictionaries by any researcher who possesses field audio recordings. Not only do these dictionaries unite phonetic, dialectal, and etymological components, but they also allow the researcher to connect each word entry to a corresponding phonetic wordform recording processed with Praat software. Further work with uploaded words is also supported by this program.

Such software is indispensable for work with endangered languages. Standard dictionaries only provide word transcriptions for extinct languages and dialects, and there is often no way to determine how accurate the transcription is. It should be noted that mistakes in transcriptions occur quite often. An illustrative example comes from the Dictionary of Selkup Dialects, published in 2005 by V. V. Bykonya; on average, Bykonya's work lists 3 to 4 transcriptional variants for each wordform. It is no longer possible to determine which variant is correct, as almost all southern and central Selkup dialects are extinct (only 1 fluent nativespeaker of Narym dialect remains). The program we are piloting at LingvoDoc will offer both scholars and future Samoyedic people the opportunity to hear pronunciations of words in these dialects long after the final speakers are gone (an eventuality we sadly anticipate happening within the next thirty years for, e.g., Narym, whose final native speaker is over 55 years old).This project will allow researchers to verify the transcriptions in [Bykonya 2005] by comparing those transcriptions to the audio files in Praat.

The fact that every user of the dictionary will be able not only to view fixed phonetic images processed in Praat, but also to work directly with the software toverify optimal processing, will dramatically increase the validity ofthe achieved results and improve worldwide communication among researchers studying endangered languages.The availability of both the dictionaries and the software online means thatsuggestions to increase the accuracy of data processing can be easily communicated and considered via the hotline.

Each of the online Samoyedic audio-dictionaries comprises about 1500 lexemes listed with paradigmatic forms. The content of each entry is as follows:

1. **Initial form of the word,**presented in the following way: 1) dictionary form (in contemporary orthography), 2) phonological or phonetic transcription of the word, 3) audio file containing pronunciations of the word, 4) image of the audio file processed using Praat phonetic software, withall main parameters reflected (intensity, duration, frequency, and tone). Note that the option exists to proceed from this image to the Praat software and independently analyze the wordform.

2. **To every initial form will be attached:** 1) pronunciations by other speakers of the same dialect, 2) inflectional wordforms (full paradigm in some cases). Every paradigmatic form and pronunciation will be presented in the same manner as the initial wordform, i.e. with orthographic notation, transcription, audio file, image of Praat processing and possibility to work further with the audio file in the program.

3. **Every initial word will have links to etymological cognates** of the lexeme in other dictionaries created by a user or a group of users who have agreed to allow public access to their dictionaries. Pressing the “Etymology” button yields a list of etymological cognates of the word, listed in the order of their relationship proximity, withmore closely related terms listed first (for example, etymological cognates from other dialects of the same language), followed by more distant cognates. Thus, for Yamal Nenets, the first words listed will be forms from other Nenets dialects — Kanin, Gydan, and others— then forms from Enets dialects, Nganasan dialects, and, finally, from Selkup dialects (Selkup is a southern Samoyedic language, while Nenets and Enets are northern Samoyedic). In the future, when dictionaries on more Finno-Ugric, Turkic and other languages have been created and hosted by this software, words inthese dictionaries will also be linked to Nenets words.

In this way, the proposed program facilitates the creationof dictionaries in which phonetic, dialectal, and etymological aspects are united. This software is also the first of its kind to offer the option to attach the results of Praat phonetic processing to every word in the dictionary and allow further work with the word in this program.

Finally, LingvoDoc introduces the possibility to provideextensive socio-linguistic description of a dialect’s current state, attach photos and videos of speakers, recorded texts in the language, and analysis the language’s phonetic, phonologic, prosodic, and etymological systems.

4 Plans

We plan to enhance LinvoDoc with further uploads of archival materials. This work will proceed in two directions simultaneously: processing of already known sources, and searching for new sources. Judging by the positive attention that has so far been paid to LingvoDoc in the literature, it is likely that we will be ableto find new sources in the archives in Saint-Petersburg, Arkhangelsk, and the Novosibirsk Audio Archive. Within the scope of this project, we also plan to introduce the option to conduct searches of Samoyedic manuscripts and audio recordings in LingvoDoc.

At present, the following materials are known but not yet properly circulated: manuscripts of M. A. Castrén(1838-1844), books on Selkup languages by N. P. Grigorovsky

(1870-1880s), books on Nenets language translated as part of the Translation Committee's activity (1879-1910), card index of A. P. Dulzon and his followers (1960-1990s), E. A. Helimski'scard index of the Northern Samoyedic Dictionary (1970-1980s).

Future work with these and newly discovered archival materials will be organized in the following way:

1. **For discovered manuscriptsand card indices, we will try to establish their dialectal affiliation.** Then we will identifythe most characteristic and frequentlyencountered linguistic features that determine the position of the manuscript within the corresponding language. Based on our knowledge of contemporary dialects, we will determine which dialect the manuscript best conforms to (in terms of phonetic and morphological features). As we search for manuscripts, we plan to monitor information on the history of the manuscripts'creation: time, place, authors, and editors. These data will be compared with published materials about missionary activity. If we don't see sufficient information about the authors of the text, we will continue searching the archive.

   The next stage of work will depend on whether we are dealing with a card index, dictionary manuscript, or textin a Samoyedic language.

2. Texts will be processed as follows:

a) The most importanttask within the present project is to determine dialectal affiliation and special features of each text. For every feature, we will identifya representative number of examples from all the manuscripts in which it is attested. Then, the features will be classified into two groups: 1) Features that are present in all historical manuscriptsof a given language, but appear to be absent in the contemporary literary language; with a high degree of probability, such features can be considered to be graphical techniques of the period; 2) features that are attested only in some manuscripts; these will be compared with the data on contemporary dialects.

As an example of a full description of the graphical-phonetic features of a manuscript, we point to the work of M. P. Bezenova [Bezenova 2014], written as a pilot study of six manuscripts of the Udmurt language. A crucial factor in Bezenova's work is the presence vs. absence of explicit indication in each manuscript of its dialectal affiliation. Among the manuscripts studied by M. P. Bezenova, five had such an indication, and one (Gospel of John) didn't.

If there is no direct indication of a manuscript's dialectal affiliation, it is necessary to try to establish that affiliation. Dialectal affiliation of translated manuscripts can be better determined by comparing these manuscripts to texts fromthe same time period whosedialectal affiliation is already known, rather than by comparing them to contemporary dialects,since there are many special features in the translations of the Gospel that are not found in contemporary dialects.

In the case of the Udmurt language, it is, first of all, the books of the Glazov and Sarapul sub-dialects.

Further, for the chosen manuscripts (the one under investigationand the sample manuscripts against which it is being compared), a full list of features discriminating these manuscripts from the literary language should be made. Next, a listof features that differentiate the studied manuscriptfrom the sample manuscriptsshould bemade. If there is almost full agreement between the studied manuscript and one of the samples, we can conclude that the text in question belongs to the same dialect. If no sample manuscript gives a high enough percentage of coinciding features, it should be concludedthat the manuscript belongs to a separate dialect.

To determine the particular dialect in which a given manuscript is written, we must consider the contemporary dialectal data of the studied language. The sub-dialect that has the most features in common with the manuscript is likely to be its descendant. If no sub-dialect seems to meet this criterion, we postulate that the dialect of the manuscript underwentextinction or drastic change within the last 150-200 years.

After the dialectal affiliation of all the manuscripts has been determined with a certain degree of probability, we proceed to step b): Comparison of the achieved data with the data on contemporary languages and dialects, and with proto-language reconstructions. We can then make a concordance of the text and then process the concordance as a dictionary. See step 3 for details.

3. **Card indices, dictionary manuscripts (**or published but less-known missionary dictionaries), and text concordances are produced by our team (see step 2.b). Weprocess these materials using the LingvoDoc software, supplementing the dictionaries in the following way: a) If we are dealing with a card index for a dictionary with materials on several languages (as E.A.Helimski's materials on three northern Samoyedic languages) or dialects (as A.P.Dulzon's data on 20 southern and central dialects), we will need to create a dictionary for each individual language or dialect, based on the data from the index. Within each dictionary, we will need to supplementevery word with aregularly spelled analog in the orthography of a widely accepted dictionary of the language. Further, words thatare reflexes of the same lexeme mustbe grouped togetherby etymological connections; the LingvoDoc software provides for this functionality.

b) For dictionaries of a single dialect, we will create a dictionary that incorporatesorthographic analogs from the literary language, or, if there is no literary language, from the most extensive and well-known dictionary of the language.

c) All such dictionaries will be connected by etymological ties to audio dictionaries based on field recordings (see above).

d) In the final stage of this work, we hope to build a new type of etymological dictionary, where audio dictionaries (based on field data) and archival dictionaries are connected in an online mode. In addition, the system offers the possibility of constant interactive replenishment of these dictionaries by any internet user who presents his or her data to the manager of the project and justifies its value.

4. **Etymological analyze:** We plan to conductan analysis of each dictionary hosted in LingvoDoc in order todescribe in detail the history of phonetic systems development in each Samoyedic dialect. We will then compare these data with already published etymological dictionaries, historical grammars, and descriptions of the Samoyedic dialects. The comparison should allow us to producea) a proto-Selkup reconstruction (lacking at the moment),b) a Proto-Northern Samoyedic reconstruction with accurately traced development of the vocalic system in the first and second syllables and quantitative correlations of vowels, c) an improved proto-Samoyedic reconstruction based on comparison of the proto-Selkup and proto-Northern Samoyedic reconstructions and incorporating data on extinct Samoyedic languages.Parallel to this work,we will search for borrowings from Tungusic and Turkic languages, which may have a separate phonological structure.

5 LingvoDoc system overview

The LingvoDoc software system provides the following features:

1.Collaborative work on dictionaries (like Google Docs or Github do)

2.HTTP REST API for all the functionality (can be integrated with any other software).

3. Web-interface(client application) that uses REST API for its system.

4. Flexible Access Control Lists (ACL) for collaborative editing, viewing and publishing.

5. Personal contribution statistics.

6. Totally customizabledictionarystructure.

7. Multilanguage translations for dictionaries based on the same data.

8. Extensible interfaces for outerapplications.

9. Scalablearchitecture (designed to utilize cloudresources for scaling).

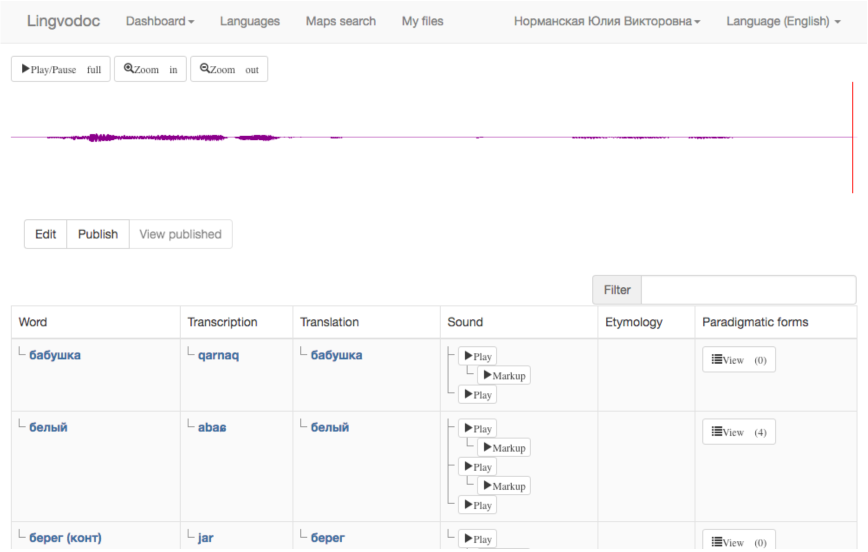
10. Semi-offline clients with 2-way sync. A user can go to the mountains or even to Mars and still sync his data, provided in internet network is accessible!

11. Multitenancy. The system natively supports total access isolation among dictionary contributors: a single user can access separate dictionaries for personal use, collaborative work, and internal usage, with each dictionary hosted at one’s own institution and shared with other set of users or institutions at will.

12. Security. We do not know users' passwords; the system is designed to hold data using the most up-to-date techniques to make sure users' data is secure.

Sources are available for public review (the license is not permissive yet [but we are planning to migrate to Apache 2.0 as soon as we are ready]; sources are providedfor informational purposes only for now).

5.2 LingvoDoc system architecture



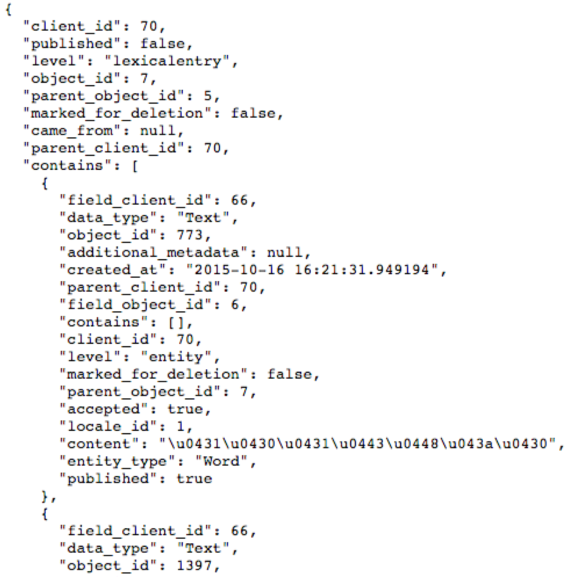
Word	Transcription	Translation	Sound	Etymology	Paradigmatic forms
бабушка	qarnaq	бабушка	▶Play ▶Markup ▶Play		View (0)
белый	abaʙ	белый	▶Play ▶Markup ▶Play ▶Markup ▶Play		View (4)
берег (конт)	jar	берег	▶Play		View (0)

Pic 1: Lingvodoc web-interface

The main feature of the system is native support for semi-offline synchronization of user data. That feature is quite unique for this kind of system and is based on the concept of a composite primary key [Date2008]. The main idea behind this kind of synchronization is that each user on each login (including offline client installations) acquires a client-unique big integer composite key. After that, each object the particular client creates is enumerated based on a special sequence identifier. On each synchronization process, the offline client acquires a new unique personal identity composite key. Thus, each object in the system has an object-unique combination of client identifier key and object identifier key. This technique allows us to make use of the "anytime synchronization" concept: for each particular offline application or client online login process, a unique object identification keyis generated.

The second basic concept is a virtual entity that doesn't contain any data in it but is an anchor to be referenced by other objects. It's quite easy to understand: imagine that you have some concept from the real world that is universal for the particular dialect you are trying to describe. In the LingvoDoc system, each global concept has a unique ID combination in composite keys terms. Each author that has an access key corresponding to that concept has the ability to add a typed entity for each global concept. That entity might be any kind the author wants: it may be a transcription, translation, media-data of any kind, a tag that will group together entities of that kind, an external link to any resource (e.g. Wikipedia), and image from the manuscript, etc. Each author can have as many versions as he wants to; the system places no limits.

In terms of data model that means that we store “versions” data in denormalized relational form and combine the data server-side. Single lexical entry example as it’s returned using our REST API:



```
{
  "client_id": 70,
  "published": false,
  "level": "lexicalentry",
  "object_id": 7,
  "parent_object_id": 5,
  "marked_for_deletion": false,
  "came_from": null,
  "parent_client_id": 70,
  "contains": [
    {
      "field_client_id": 66,
      "data_type": "Text",
      "object_id": 773,
      "additional_metadata": null,
      "created_at": "2015-10-16 16:21:31.949194",
      "parent_client_id": 70,
      "field_object_id": 6,
      "contains": [],
      "client_id": 70,
      "level": "entity",
      "marked_for_deletion": false,
      "parent_object_id": 7,
      "accepted": true,
      "locale_id": 1,
      "content": "\u0431\u0430\u0431\u0443\u0448\u043a\u0430",
      "entity_type": "Word",
      "published": true
    },
    {
      "field_client_id": 66,
      "data_type": "Text",
      "object_id": 1397,
```

Pic 2: Single lexical entry representation example

Let’s imagine, for instance, that M authors have different opinions on some object, such as the translation of a particular term. The system doesn’t limit the number of listed translations, provided each of the M authors has corresponding rights for the dictionary.

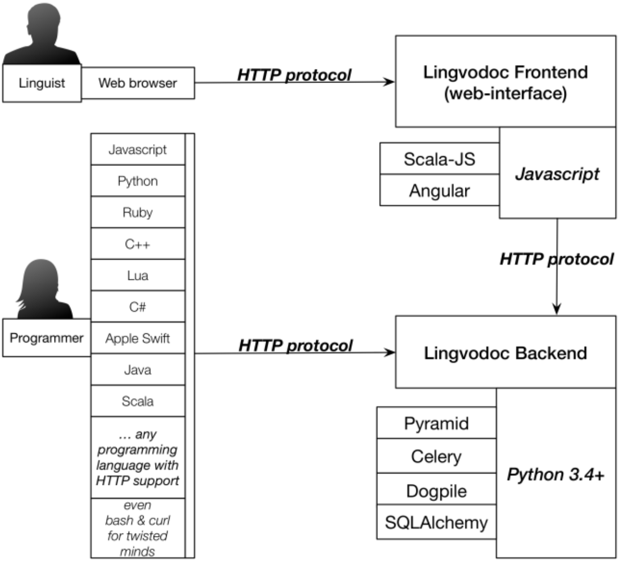
But! The system provides special views for that purpose. The main view, of course, is the editor’s view: from there, the authors of the data can do anything they want. The second view is the publisher’ s view. Using that view, the people responsible for a dictionary can approve one or more correct entities used in the virtual anchor object. For instance, imagine that some lexical entry has 5 versions for transcriptions and 10 versions for translations. Let’s suppose that the owner of this dictionary thinks that only one of the transcriptions and three of the translations are correct. He can select only them and publish his choices to other researchers!

The last view is the view/guest/data-researcher’ s view. Here you can see the data that have been uploaded and verified by authors and publishers.

5.3 LingvoDoc system outer access

Of course, LingvoDoc offers standard web-interface access, but that is not too interesting from the perspective of concurrent technologies. **Full access to the LingvoDoc system**

using can be gained through the REST API (HTTP-based) system. Each object in the system has a clear access method; our web-interface is just a reference javascript client. All levels of access are available using our simple HTTP-based API. The system arch looks like this from the user's point of view:



Pic 3: Lingvodoc architecture overview

You can gain any kind of access to the system provided that you have been granted author permission.

Each author of any dictionary has a right to distribute access to his data for any particular user or organization in the system.

Possible use-cases for the LingvoDoc system:

1. Data analytics in your programs (using REST API).
2. Media- and etymology- based dictionaries with audio markup.
3. Images with corresponding texts for images.
4. Audio data with comments for each audio segment.
5. Much more, as soon as we have freed ourselves from restrictions ondata connections.

Finally, as noted above, the system is able to show users markup for any sound in Praat or ELAN format. This functionality will soon be extended to include external functions for corpora enrichment (e.g. parser extensions).

Literatur

Bezenova (2014). Phonetic features of manuscripts of the Udmurt writing of religious character of the first half of the XIX century. In Ural-Altaic Studies. 1. 2014, pages 22-44.

Bykonya (2005). Sel'kupsko-russkij dialektnyj slovarj pod redakciej V.V.Bykonya. Tomsk, 2005.

Date, C. (2008). The relational database dictionary. Apress, page 32.

Helimski (2000). Comparativistik. Uralistic. Lekzii i stat'ji. Moskva, 2000.

Janhunen (1977). Samojedischer Wortschatz. Gemeinesamojedische Etymologien. Helsinki, 1977.

Mikola (2004).Studien zur Geschichte der samojedischen Sprachen. Szeged, 2004.