МНОГОФУНКЦИОНАЛЬНЫЙ ИНТЕРНЕТ-РЕСУРС «МАЛЫЕ ЯЗЫКИ СИБИРИ: НАШЕ КУЛЬТУРНОЕ НАСЛЕДИЕ»

Казакевич О. А. (kazakevich.olga@gmail.com),
Воронцова М. И. (marinavorontsova@yandex.ru)
НИВЦ МГУ им. М. В. Ломоносова, Москва, Россия

Клячко Е. Л. (elenaklyachko@gmail.com)
НИУ Высшая школа экономики, Москва, Россия

Поливанов К. К. (polivanov.studio@gmail.com)
Факультет журналистики МГУ
им. М. В. Ломоносова, Москва, Россия

В докладе предполагается рассказать о результатах первого этапа работы над проектом «Создание интернет-ресурса „Малые языки Сибири: наше культурное наследие“ (на материале языков бассейна Среднего Енисея и Среднего и Верхнего Таза)» (http://siberian-lang.srcc.msu.ru). Проект реализуется в лаборатории автоматизированных лексикографических систем Научно-исследовательского вычислительного центра МГУ им. М. В. Ломоносова при поддержке РГНФ, грант 12-04-12049в. Основой информационного наполнения ресурса (сайта) служит мультимедийный компьютерный архив материалов по говорам исчезающих языков региона — прежде всего, селькупского, кетского и эвенкийского, — записанных в ходе регулярно проводившихся на базе лаборатории в течение последних полутора десятков лет лингвистических экспедиций. Кроме того, при работе над сайтом используются изданные и архивные материалы по этим языкам, отражающие разные временные срезы их функционирования на протяжении последнего столетия.

Разработчики сайта ставят перед собой две задачи — научную и просветительскую. С одной стороны, сайт адресован научному сообществу, прежде всего лингвистам, антропологам, историкам, с другой — носителям языков, представленных на нем, а также тем, кто эти языки преподает или изучает, хочет их выучить или просто хочет что-то узнать о языке и культуре сибирских народов, важной части культурного наследия нашей страны. Процесс языкового сдвига, разные стадии которого фиксируется сегодня во всех этно-локальных группах селькупов, кетов и эвенков, в немалой степени связан с низким статусом этих языков, прежде всего в молодежной среде. Неоднократно отмечалось, что присутствие языка в Интернете повышает его привлекательность для молодежи, поэтому создаваемый сайт рассматривается разработчиками и как средство поддержки языков, материалы которых на нем размещаются.

На сайте представляются данные следующих типов: 1) социолингвистические данные, организованные в базу и характеризующие сохранность языков в обследованных в ходе экспедиций поселках;

2) мультимедийные словарные базы данных селькупских, кетских и эвенкийских говоров региона; 3) аннотированные тексты на селькупских, кетских и эвенкийских говорах как записанные в ходе экспедиций последних лет, так и взятые из архивов; большинство текстов современной записи имеют визуальное, звуковое и графическое представление, состоящее из текста в официально принятой для каждого из языков графике, фонетической транскрипции, отражающей специфику каждого из говоров соответствующего языка, и пофразового перевода текста на русский язык; архивные тексты представляются только графически в исходной графике собирателя, в современной официально принятой графике и в фонетической транскрипции; разметка текстов состоит из характеристик, приписываемых каждому тексту как целостной единице, и внутренней разметки каждого из текстов — поморфемной грамматической индексации; 4) тексты на русском языке, записанные в экспедициях от русского старожильческого населения и рассказывающие о жизни, обычаях и ритуалах селькупских, кетских и эвенкийских соседей; 5) данные о грамматике каждого из трех языков в виде грамматических справочников; 6) обучающие программы для каждого из языков; 7) документальные фильмы о функционировании языков и о жизни их носителей; 8) фотоматериалы, отражающие повседневную жизнь обследованных поселков.

Ключевые слова: многофункциональный интернет-сайт, исчезающие языки Сибири, мультимедийные аннотированные корпусы текстов, мультимедийные словари, наше культурное наследие

MULTI-FUNCTIONAL WEB-SITE “MINORITY LANGUAGES OF SIBERIA AS OUR CULTURAL HERITAGE”

Kazakevich O. A. (kazakevich.olga@gmail.com),
Vorontsova M. I. (marinavorontsova@yandex.ru)

Research Computing Centre, Lomonosov Moscow State University, Moscow, Russia

Kliachko E. L. (elenaklyachko@gmail.com)

Higher School of Economics, National Research University, Moscow, Russia

Polivanov K. K. (polivanov.studio@gmail.com)

Journalistic Faculty, Lomonosov Moscow State University, Moscow, Russia

The paper presents some results of the ongoing project “Development of the web-site ‘Minority languages of Siberia as our cultural heritage’ (on the material of the languages of the basin Middle Yenisei and the Middle and the Upper Taz)” (http://siberian-lang.srcc.msu.ru). The project is being realized at the Laboratory for Computational Lexicography, Research Computing Centre, Lomonosov Moscow State University, with financial support from Russian Foundation for the Humanities, grant 12-04-12049v. The primarily informational source for the site is the Multimedia Computer Archive of materials in local dialects of Selkup, Ket and Evenki, recorded in the course of linguistic expeditions regularly organized and led within the laboratory, but we are going to draw relevant information from published works and archives as well.

Working at the site, we keep two target groups in focus: on the one hand, we addressed the academic community — linguists, anthropologists, historians, on the other hand, we addressed the speakers of the languages presented on the site, those who teach or learn these languages at schools and colleges, who want to learn them or just to get acquainted with languages and cultures of Siberian peoples, an integral part of our cultural heritage. The project will result in a multi-media web-site containing information on the present-day situation of three endangered minority languages of Siberia, multi-media lexical databases and annotated text corpora in local varieties of these languages, as well as language learning materials and software, which can help to acquire Selkup, Ket or Evenki grammar and lexicon.

Key words: multi-functional web-site, endangered languages of Siberia, multi-media text corpora, multi-media dictionaries, our cultural heritage

0. Introduction

Documentation and description of endangered languages of Siberia is one of the research branches of Laboratory for Computational Lexicography of Research Computer Centre, Lomonosov Moscow State University. This branch has been developed in the Laboratory since 1987. Since then a series of projects were realized, which were aimed at documentation and description of endangered languages of Siberia using modern technologies of audio and video recording, as well as at creation of computerized multimedia archives for preserving the collected field data and for providing a convenient access to the data for their future processing. At present the languages we are working with — Selkup, Ket and Evenki, — are all endangered. All the Selkups, Kets and Evenkis are fluent in Russian and for many of them Russian is the only language they speak: in all Selkup, Ket and Evenki communities language shift process is at hand. It cannot be said that the problem of heritage language attrition does not worry the communities. The attitude of the ethnic community members towards the heritage language is mostly positive, the majority of parents say they want their children to speak their ancestral language, but even those parents who are able to speak the ancestral language themselves choose to speak Russian with their children. Thus, though the ancestral language is regarded as desirable for children it is Russian that is considered obligatory. Actually, people do not believe that anything can be done to stop the shrinking of the use of their ancestral languages. Many of them just state

their heritage language is dying. Meanwhile, those dying languages are not only the cultural heritage of Selkup, Ket, or Evenki ethnic communities, they are an integral part of our national heritage, and it does not seem wise just to let this part be perished. Can the situation be somehow reversed? Today it is agreed among the experts that the presence of a language in the Internet elevates its prestige and attractiveness for the younger generation (see e.g. (Kazakevich 2007)). This argument was among some others which led the authors to the idea to develop a multi-functional web-resource, which could be interesting and useful both for academic communities of linguists, anthropologists and historians, and for ethnic communities of Selkups, Kets and Evenkis, as well as for all those who just want to know about autochthonous peoples and languages of Siberia. As the primarily informational source for the site we use the multimedia computer archives of linguistic materials in local dialects of Selkup, Ket and Evenki, recorded in the course of almost two dozens of linguistic expeditions organized and led on the basis of the laboratory, but we are also going to draw relevant information from published works and archives. On our site, we try not only to give linguistic data but also to represent sociolinguistic and cultural context in which these data were recorded. The project design was fine, but only having started the project after we received a grant for its realization, we fully understood what tremendous amount of work we are exposed to.

1. The overall design of the website

The objective of the project is the development of a multi-media website containing comprehensive data on autochthonous minority languages of the basins of the Middle-Yenisei and the Middle- and Upper-Taz — Selkup, Ket and Evenki. On the site the following types of data are supposed to be found: 1) sociolinguistic data characterizing the functioning and degree of preservation of Selkup, Ket, or Evenki in the surveyed villages; 2) sounding thematic dictionaries of Selkup, Ket and Evenki local dialects recorded in the field within the last decade and a half and loaded into a database, which is to be developed for that purpose; 3) annotated Selkup, Ket and Evenki text corpora containing texts of various genres in local dialects of these languages both recorded in the course of our expeditions of the last decade and a half and extracted from archives (first of all, from the archive of Peter the Great Museum of Anthropology and Ethnography, Russian Academy of Siences, Saint Petersburg; each recently recorded text will be represented with an audio file, with a graphic version in the officially adopted for each language graphic system, with a phonetic transcription approaching a phonemic one, but reflecting specific features of each local dialect, and with a phrase-for-phrase Russian translation; some texts will be also represented with video files; archival texts will be represented only with graphic files in the graphics used by the researcher, who had recorded it, in the modern officially adopted graphic system and in phonetic transcription; the annotation will consist of characteristics attributed to each text as a whole and characteristics assigned to different units inside the texts; some texts will be supplied with morphological indexing, they will be represented in the ELAN format; 4) Russian texts recorded in the field from the Russian

population of the area living side by side with the Selkups, Kets and Evenkis and telling about life, traditional ways and rituals of their neighbous; 5) information on the grammar structure of the three languages in the form of grammar reference book or reference database; 6) Selkup, Ket and Evenki language learning software; 7) documentaries showing the functioning of the languages and the life of their speakers; 8) photos representing the life of the villages and their residents, our informants and the nature surroundings.

2. The first year work at the project

The data model for the new site was our first step. We adopted it having analyzed the material we disposed of and precised some details of the site conception, taking into account the experience gained from the work at the earlier created demo-versions of Selkup, Ket and Evenki lexical databases (Kazakevich et al 2004; 2005; 2007). After long discussions the research group finally developed the first version of the content format for different types of materials such as accessibility of basic sociolinguistic data to ‘inner’ and ‘outer’ users, or how ‘open’ the information of the language competence of our informants should be, as well as the relevance of two types of *places — place of residence* and *place of data collection*. Then we passed to the site-design.

2.1. Website design.

The guiding principal in designing the website was to make it accessible and usable by a wide spectrum of users. Neither professional linguists doing a research on the comparative analysis of phonetic features, nor people living in a faraway village by the River Taz or Turukhan interested in Ket folk songs or Evenki life stories, or looking for photos of their home town, would feel uncomfortable using the website.

Thus, the Web-Resource should meet the following criteria:

- it is easy to use
- it is well-structured
- it is easy to find relevant information
- the content is in an easy to use format
- it provides facilities to help locate information
- it has a search facility, a logical navigation menu, a site map or an index
- it loads quickly
- it is attractive in design
- the content is copyright or it can be used providing the source is acknowledged
- it is technically stable

The further work at the project consisted of two equally important parts: the web-site and databases construction, on the one hand, and data processing to supply informational content for the site and its databases, on the other hand.

2.2. Site construction

2.2.1. Basic details

To develop user interfaces and an interface to edit and create materials we used Javascript, CSS3, html5 (Hogan 2012; McCaw 2012; Simpson & Schmitt 2012). As a content management system we chose Drupal (Tomlinson 2011), a free and open-source content management framework (CMF) written in PHP and used to store open source relational database management system MySQL (Davis & Phillips 2008).

2.2.2. Preliminary work on the server

Our website is hosted by the server of Research Computing Centre, Lomonosov Moscow State University. The integration into the MSU website system ensures stable functioning of our website, which we find most important. To display the designed site on the server some preparatory work had to be done. It consisted of

1) generating a key to access the files on the server SSH;
2) participation in the actual installation of the server version of relational database MySQL;
3) installing the system-date version of phpMyAdmin, which is designed to work with databases via web-interface;
4) setting up a MySQL user and database creation;
5) participation in setting up and configuring the current version of PHP, one of the most popular scripting languages in the programming for the Internet;
6) participation in installing apache-server fnd setting up apache.

2.2.3. On the technical means

While developing the website, we use the latest relevant hardware and software, which enables cross-browser compatibility and cross-platform operation of the site. The site is to be displayed properly on different devices including netbooks, tablets and other mobile devices. Cross-browser refers to the ability of a website to look and run *identically* in all popular browsers, which means the absence of the dismantling layout and the ability to display data with the same degree of readability.

2.2.4. Core of the site

As basic core software we chose the latest version of Drupal — Drupal 7, a content management system (CMS) written in PHP and used as a data warehouse relational database (supported by MySQL). Drupal is open source software distributed under the terms of GNU General Public License (or “GPL”). It is constantly developed by an active and diverse community of people around the world. Drupal meets all modern standards including high level of security. Drupal was used to build the website of Russian Festival of Science, the White House in the USA and many others.

2.2.5. Design and Programming

Designing and programming the site we had to fulfill the following tasks:

1) to design the logical structure of the site, based on a detailed analysis of the materials;

2) to find the most convenient solutions of information presentation;
3) to create software templates for filling the site databases, and for that purpose to fully describe the their structure and information;
4) to programme and configure the system dictionary of terms (taxonomy), which should be related to the material; classifying the data category, taxonomy provides a classification and ordering of materials;
5) to programme search engine, based on categories and keywords;
6) to programme a player that plays online videos and audio files;
7) to cascade Style Sheets (CSS);
8) to make-up HTML-pages;
9) to programme the account and the corresponding user roles;
10) to programme content management system, which could be reliable and intuitive.

2.2.6. Interface

While developing the site we use the latest technology HTML5 and CSS3, which should provide an elegant, comfortable and modern interface. Besides, adapting materials provided for viewing on older systems, so that people in the depths of Siberia could easily use materials of the site and the database. In addition, the interface of the website is constructed by means of innovative solutions built on the base jquery (library of JavaScript, which focuses on the interaction of JavaScript and HTML). It also provides for the possibility of publishing in the popular social networks without leaving the site.

2.3. Adding dictionaries, texts, and sociolinguistic data

2.3.1. Adding «inherited» data

An SQL-script which inserts the «inherited» dictionary data (previously stored in Excel spreadsheets) into the database has been written. There have also been created VBA-scripts which import MSWord documents into ELAN projects (ELAN — Linguistic Annotator).

2.3.2. Adding new data

New dictionary data (audio files and transcriptions of Evenki dictionaries) has been uploaded. Several texts have also been uploaded with links to the same texts on Languedoc project provided (Languedoc project). A program to automatically use SIL converters (SILConverters 3.1.1 Overview) for adding the Cyrillic version of Evenki texts has been written. We use the Languedoc server as the main storage place for the texts, as the TLA tools (TLA tools) provide a lot of services which are necessary for analyzing the texts. Firstly, the TLA tools are integrated with the Fieldworks Language Explorer tool (Language Explorer (FLEx)), which we use for interlinearising the texts. Secondly, the platform enables a viewer to search the annotated text, the search options including metadata as well as annotation data fields, and view it using a user-friendly interface. Thirdly, the multilevel annotations created with ELAN can be exported in the standard SRT subtitle format.

2.4. Testing the work of the site functionality

The functionality of the site was tested manually as far ad data adding and site viewing concerns.

References

1. *Hogan B.* (2012) HTML5 and CSS3. Web Development by the standards of the new generation. St. Petersburg: Peter.
2. *McCaw A. (2012) Web-based applications to JavaScript. St. Petersburg: Peter.*
3. *Simpson K., Schmitt K.* (2012) HTML5. Recipes programming. St. Petersburg: Peter.
4. *Tomlinson, Todd* (2011). Pro Drupal development (third edition), apress. Moscow: Publishing House „Williams".
5. *Davis, Michele E. and Jon A. Phillips* (2008) Learning PHP and MySQL (second edition). O'railly Media. St. Petersburg: Simvol Plus.
6. *Kazakevich O. A.* (2008) Supporting minority languages with the help of the informational-communicative technologies: international experience [Podderzhka malykh yazykov s pomoshshiu informatsionno-kommunikatsionnykh tekhnologii: zarubezhnyi opyt] // Predstavleniye yazykov Rossii i stran SNG v rossiiskom segmente Interneta [Presentation of the languages of Russia and of the CIS in the Russian Internet segment]. Seminar Rossiiskogo komiteta Programmy UNESCO "Informatsiya dlia vsekh" I Mezhdunarodnoi konferentsii "EVA 2007 Moskva". [Seminar of the Russian Committee of UNESCO Programme "Information for all", Ist International Conference "EVA 2007 Moscow"]. Moscow: International Library Centre. Pp. 15–26.
7. *Kazakevich O. A., Zakharov L. M., Samarina I. V., Trushkov D. L.* (2004) Corpus linguistics, computer lexicography, multimedia technologies, and endangered languages (The project is completed — Long live a new project!) [Korpusnaya lingvistika, kompiuternaya leksikografiya, mul'timediinyie tekhnologii i ischezayushiye yazyki (Proyekt zavershen — da zdravstvuyet novyi proyekt!)] // Computational linguistics and intellectual technologies. Proceedings of the International Conference Dialogue'2004. Moscow. Pp. 252–257.
8. *Kazakevich O. A., Samarina I. V., Itkin I. L., Bagariatskaya T. B., Reutt T. E.* (2005) The Ket project: what has been done within a year [Ketskii proyekt: godovoi tsikl rabot i rezul'taty pervogo kruga] // Computational linguistics and intellectual technologies. Proceedings of the International Conference Dialogue'2005]. Moscow. Pp. 228–232.
9. *Kazakevich O. A., Itkin I. L., Mitrofanova N. K., Reutt T. E.* (2007) Multimedia database of Evenki local dialects and standard Evenki [Multimediinaya baza dannykh evenkiyskikh govorov i evenkiyskiy literaturnyi yazyk] // Voprosy filologii [Journal of Phylology]. N 2 (26). Moscow. Pp. 42–47.