Language Resources and Evaluation

The software system LingvoDoc and the possibilities it offers for the documentation and analysis of Ob-Ugric languages

--Manuscript Draft--

Manuscript Number:	
Full Title:	The software system LingvoDoc and the possibilities it offers for the documentation and analysis of Ob-Ugric languages
Article Type:	Full length article, original research
Keywords:	Ob-Ugric languages; LingvoDoc; Collaborative dictionary system; REST API; Sound processing; speech processing; Praat; ELAN
Corresponding Author:	Oleg Borisenko Institute for System Programming of the Russian Academy of Sciences Moscow, RUSSIAN FEDERATION
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Institute for System Programming of the Russian Academy of Sciences
Corresponding Author's Secondary Institution:	
First Author:	Julia Normanskaja, Prof. PhD.
First Author Secondary Information:	
Order of Authors:	Julia Normanskaja, Prof. PhD. Oleg Borisenko Ivan Beloborodov
Order of Authors Secondary Information:	
Funding Information:	Russian Federation government (14.Y26.31.0014) — Not applicable
Abstract:	LingvoDoc system (http://lingvodoc.ispras.ru) provides a service for collaborative language documentation and computations on the collected data. This software system provides HTTP REST API for all the system components and allows to build own extensions for data analysis or even to integrate with your own software. Thanks to a special database and application design pattern it's possible to construct offline applications integrated with Lingvodoc system: these applications would need to have the internet only once to synchronize basic datatypes and for authentification purposes. The system itself allows users to construct multilayer dictionaries, attach them to map, populate with metadata documents, share access to dictionaries with other users or with everyone. The LingvoDoc system provides fine-grained access control lists for sharing which allow to have separate groups of users for dictionary edits, for proofreaders and for read-only users. The system also provides some computational algorithms on the stored data: phonology computations, automatic and guided deduplication inside the dictionaries etc. The system allows users to choose dictionary structure. The system supports the following datatypes: text, images, sounds (wav and mp3), markups (ELAN and Praat formats), directed and undirected links between stored entities. User creates the most suitable for his dictionary. Also the system provides ELAN corpora storage, viewer and processing. This paper shows documentation and analysis of Ob-Ugric languages using LingvoDoc system.

The software system LingvoDoc and the possibilities it offers for the documentation and analysis of Ob-Ugric languages.

Abstract. LingvoDoc system (http://lingvodoc.ispras.ru) provides a service for collaborative language documentation and computations on the collected data. This software system provides HTTP REST API for all the system components and allows to build own extensions for data analysis or even to integrate with your own software. Thanks to a special database and application design pattern it’s possible to construct offline applications integrated with Lingvodoc system: these applications would need to have the internet only once to synchronize basic datatypes and for authentification purposes. The system itself allows users to construct multilayer dictionaries, attach them to map, populate with metadata documents, share access to dictionaries with other users or with everyone. The LingvoDoc system provides fine-grained access control lists for sharing which allow to have separate groups of users for dictionary edits, for proofreaders and for read-only users. The system also provides some computational algorithms on the stored data: phonology computations, automatic and guided deduplication inside the dictionaries etc. The system allows users to choose dictionary structure. The system supports the following datatypes: text, images, sounds (wav and mp3), markups (ELAN and Praat formats), directed and undirected links between stored entities. User creates the most suitable for his dictionary. Also the system provides ELAN corpora storage, viewer and processing. This paper shows documentation and analysis of Ob-Ugric languages using LingvoDoc system.

Funding. This report has been published within the framework of the project, “Language and ethno-cultural variability of Southern Siberia in synchrony and diachrony: language and culture interaction” (Russian Federation government grant No. 14.Y26.31.0014).

At the beginning of the 20th century, the Ob-Ugrians — speakers of Khanty and Mansi — still occupied a vast territory extending from the upper reaches of Pechora, in the northern Urals, to the Yugan, Vasyugan, and Vakh rivers in the Tomsk district (a total land area of about 3000 km, northwest to southeast). It’s not surprising that related languages distributed over such a large area should show significant dialectal disunity. Khanty and Mansi are divided into four dialectal groups each, and there is no mutual understanding between speakers of these groups. At the beginning of the 20th century, each of these groups contained several dialects, with significantly different morphological and phonetic systems. Even within a single dialect, sub-dialects significantly varied. Sos’va Mansi, a dialect spoken in the Sos’va and Lombovozh villages about 1000 km apart, provides a good illustration. Our latest studies show that the Sos'va sub-dialect of Sos'va Mansi contains both long and short vowels (as is the case in literary Mansi), while a phonetic software analysis of data collected by Novosibirsk researchers in the 1970s shows that the Lombovozh sub-dialect, spoken to the south in the Sverdlovsk district, had only neutral vowels. All told, the dialectal differences among the Khanty and Mansi dialectal groups in the beginning of the 20th century permit division into eight languages and tens of dialects, all differing significantly from each other.

Sadly, this situation is now changing catastrophically quickly. Some dialectal groups have already gone: the last speakers of Southern and Eastern Mansi died in the middle of the 20th century, as did the last speakers of Southern Khanty. Only one speaker of an Eastern Mansi dialect remains alive today (see below). Some Khanty dialects, like Nizyam (a transitional group between Southern and Northern Khanty) and Salym (a transitional group between Western and Eastern Khanty), have been considered dead, but field expeditions conducted by researchers within our project (S. Onina and M. Amelina; see below) have located a few remaining speakers.

According to the 2010 census, there are about 9,500 remaining Khanty speakers. These speakers almost exclusively speak northwestern dialects; as noted above, the southwestern dialects have already disappeared, and the eastern ones are only spoken by a few elderly individuals (70–80 years old) who live in villages at the Ob' tributaries. Eastern Khanty dialects are not used in daily life.

According to the same census, about 4,500 Mansi speakers remain, speaking almost exclusively northern dialects (Sos'va and Sygva). There are no more Western or Southern Mansi speakers, and we know of only one Eastern Mansi speaker.

In the last few years, linguists from different countries have undertaken important work to record and study the Ob-Ugric languages, realizing their critical condition. The two largest of these projects are:

- EuroBABEL Ob-Ugric Languages: an Ob-Ugric database of analyzed text corpora and dictionaries for less-described Ob-Ugric dialects, led by E. Skribnik at http://www.babel.gwi.uni-muenchen.de/; also at that address are housed detailed links to numerous resources on Ob-Ugric languages and ethnography. Field data on the Kazym, Surgut and Yugan Khanty dialects, together with data on the Sos'va Mansi dialect, were collected through this project, as well as glossed texts of northern, western and eastern Mansi dialects and northwestern and eastern Khanty dialects.
- Multimedia documentation of the endangered Vasyugan and Alexandrovo Khanty dialects of the Tomsk region of Siberia, led by A. Filchenko at

http://www.policy.hu/filtchenko/FTG%20ELDP%20project/audio.htm. This project collects and analyzes Eastern Khanty texts.

Despite these important projects, before our work started, there was no audio data accessible to the public on the Eastern Mansi dialect, the Ob' sub-dialect of Northern Mansi, or the Nizyam and Salym transitional dialects. Furthermore, numerous Khanty and Mansi texts, created in Russia in the 19th century, had never been analyzed. Our project has uncovered dozens of books in the archives and libraries of Saint Petersburg and Finland, including gospels, liturgical texts, and various dictionaries of partly disappeared dialects of Khanty and Mansi.

Since 2012, our group has been conducting research to identify accent systems with moving stress in Ob-Ugric languages. We have organized a number of expeditions into remote regions of western Siberia and, with the help of local administration, located numerous Khanty and Mansi speakers.

In 2012, S.V. Onina led an expedition to visit the last remaining Khanty speakers, who live in the woods along the river Nazim; the expedition identified them as speakers of the **Nizyam dialect**, previously considered extinct. A complete dictionary of the Nizyam dialect, containing more than 2,500 lexemes, was compiled with the help of these informants. In her analysis of this material, Ju.V. Normanskaja reported the existence of free phonologically significant stress in non-derivative names; the placement of this stress corresponds to the position of stress in southern Irtysh dialects, according to data discovered by Ju.V. Normanskaja in the archives of M.A. Castren. (For further details, see: Ju.V. Normanskaja (2014), “Stress system in Nizyam Dialect Of Khanty Language (based on field data) and its parallels in South Khanty (based on materials from M.A. Castren’s archive [Normankaja 2014]).

In 2013, M.K. Amelina, with the help of E.V. Korovina, conducted field research with speakers of the Salym and Surgut dialects of Khanty and Yukonda Mansi. The Salym dialect was previously considered extinct; however, during the 2013 expedition, a whole village was found in which Salym is the main language of everyday communication, spoken also by children. A dictionary survey was conducted and is currently being processed.

Field work with native speakers of the **Yukonda dialect of Eastern Mansi** also took place in 2013, with material collected by M.K. Amelina in Shugur village in the Kondinsk area, Khanty-Mansi Autonomous Okrug. At the time of this expedition, only two native speakers of the language were found: Selivanova (Nohova) Elizaveta (Anna) Ivanovna, 89 years old, and

Shivtorov Maksim Semenovich, 74 years old; other speakers remember only a few words. Full vocabularies of native lexis with paradigmatic forms and usage contexts were recorded from the two speakers, as well as short texts. Altogether, about 700 lexemes, each with a fair number of wordforms (5 to 15 forms for each lexeme) were recorded. Soon after the departure of the expedition, Selivanova (Nohova) Elizaveta (Anna) Ivanovna died. The recordings made by M.K. Amelina on this trip are thus particularly valuable, since there now remains only one speaker of Eastern Mansi. Recordings from Selivanova (Nohova) Elizaveta (Anna) Ivanovna confirm that the features present in Maksim Shivtorov's speech are not idiolectal. Ju.V.Normanskaja's analysis of the collected material using Praat phonetic software revealed a free paradigmatic accent in the East Mansi verb system that systemically corresponds to the placement of stress in the Ob dialect of North Mansi. (For further details, see [Normanskaja 2015b]).

I.A. Stenin conducted field work with speakers of the **Ob Mansi dialect** living in two villages of the Oktyabrsky region of KhMAO, Nizhnie Narynkary and Peregryobnoe. He polled four native speakers of Ob Mansi from Nizhnie Narynkary (Butarina Galina Vasilyevna, Kulikova Nina Iosifovna, Matveeva Zinaida Yakovlevna, and Plekhanova Klavdiya Semyonovna), and two speakers from Peregryobnoe (Gyndysheva Taisiya Grigoryevna and Yarlina Evdokiya Grigoryevna). Where possible, Stenin recorded word lists of native vocabulary. Almost all speakers of the Ob dialect are older than 60. In childhood, these speakers spoke only Mansi; they didn't learn Russian until they went to school. At present, they speak only Russian with their children and grandchildren, while among each other they speak Mansi but often switch to Russian. Phonetic differences between of the Ob dialect and literary Mansi are summarized in [Rombandeeva, Kuzakova 1973; Sajnahova 2012], but the authors emphasize that the Ob dialect as a whole needs further study. A general lack of detailed research may explain the fact that previous scholars haven't noticed the moving accent in the Ob verbal system. A full Praat analysis of the collected material, segmented into word-forms, showed evidence for three verbal accentuation paradigms in the Ob dialect of Mansi. (For further details, see: [Normanskaja 2015a]).

A major component of our project, in addition to field collection of material on the Ob-Ugric languages, involves analyzing that data using Praat phonetic software and identifying etymological connections between the various dialectal materials. It is now possible to do this work online at our project website: http://lingvodoc.ispras.ru/ (for more information, see the second part of this article, below).

This analysis has allowed us to identify new patterns of phonologically significant free stress in some dialects of Finno-Ugric; these patterns were unknown to 20th century Ob-Ugric scholars.

In 2011, our work with the archives of M.A. Castrén's work (in the National Library of Finland in Helsinki), uncovered Castrén's records of a similar moving accent pattern in the Southern Khanty dialects of the Irtysh area. This is very valuable data, since these Southern Khanty dialects don't exist anymore. Castrén's word list introduces about two hundred words, mostly non-derived nouns and verbs in the first singular present or preterit plus infinitive. For all Castrén's recorded infinitives, the stress falls on the first syllable, while in oblique cases and in the first singular Present time or Past time verb forms, stress can fall on either the first or second syllable. We could not detect any correlation between vowel quality and stress placement. These data lead us to hypothesize that non-derivative words in Southern Khanty had paradigmatic phonologically important stress.

To date, we have identified moving stress in several Khanty and Mansi dialects, including Salym[1] and Surgut (Eastern Khanty), Nizjam (halfway between Northern and Southwestern Khanty), and Yukonda and Ob (Mansi). We recorded basic word lists of non-borrowed Finno-Ugric vocabulary for each dialect using data from two or more native speakers[2]. Our group is currently processing the received data; beta versions of the etymological audio-dictionaries of these dialects can be accessed at http://lingvodoc.ispras.ru/. We are also continuing to study Castrén's archival data on the Khanty dialects, where moving stress was attested.

The first interesting results retrieved during our description and comparison of the individual accent systems were published in [Normanskaja 2013, 2014]. In [Normanskaja 2013], we presented the results of our analysis of stress placement in Vasyugan, based on archival data collected by L.I. Kalinina in the 1950 and 1960s. We showed that the rules of stress placement depend on the part of speech of the word. For verbs and pronouns, stress is paradigmatic, and its placement depends on the type of affix. For nouns, the stress is fixed on either the first or second syllable.

Fixation of the stress on a certain vowel seems to have occurred long ago, at a time when first-syllable vowels still retained their proto-Khanty quality in the eastern dialects. If the first-syllable vowel was high, then the stress became fixed on the second syllable; otherwise, it became fixed on the first syllable. Thus, the Vasyugan accent system turns out to be unique within the Khanty family in exhibiting nouns and verbs with distinct stress patterns: verbal

[1] Reports from previous researchers indicated this dialect was extinct; cf. [Nikolaeva 1999]. M.K. Amelina has found several dozen people who have good command of the language, but only one woman among them – Kayukova Svetlana Petrovna – has agreed to cooperate.

[2] Salym is an exception, as the word-list for this dialect was recorded from a single speaker and is not complete.

stress is in a moving paradigm, while noun stress is fixed and depends on the phonemic composition of the word.

In [Normanskaja 2014], we showed that stress in non-derived words in the modern Nizjam dialect coincides completely with the Irtysh (Southern Khanty) data recorded by M.A. Castrén. In derivative stems in Nizjam, categorization of metrical stress occurred; in the modern language, in the absence of non-derivative dictionary forms, it is impossible to determine the original stress placement. The fact that we observe paradigmatic accent systems represented by non-derivative nouns and verbs in both Nizjam and Southern Khanty indicates that such a system was present in the archaic (at least, proto–Western Khanty) language.

Continuing the cycle of articles dedicated to the description of moving stress in Ob-Ugric languages, [Normanskaja 2015a, b] considers the moving stress in Mansi dialects and show that it has non-trivial parallels in the Southern Khanty data of M.A. Castrén.

Moving accent in Mansi was noted as early as the 19th century in [Munkácsi 1894]. The author lists verb paradigms for Eastern, Western, and Southern Mansi dialects, from which we can see that, for some verbs, the accent is fixed on the first syllable, while for others, the accent moves within the paradigm. Unfortunately, the materials found in the monograph are by no means exhaustive. Although her data illustrate verb paradigms for every dialect, Munkácsi gives a few lexemes for which the placement of stress is marked only sporadically in the western and, especially, eastern dialects.

For Southern Mansi, stress is marked on the first or second syllable for all forms. There are at least two verb accent paradigms: one with fixed stress on the first syllable, and one with stress on the second syllable for all present and past tense forms and on the first syllable for future tense forms. In Munkácsi (1894), the author noted moving stress in the verb paradigms of all dialect groups except the northern; however, this work offers very little data concerning which lexemes had fixed stress, and which had moving stress. Much more informative in on this topic is offered in the dictionary [Munkácsi, Kálmán 1986]. In this work, the placement of stress is marked for practically all Southern Mansi verb lexemes in the present third person singular form. From time to time, other verb forms are also provided, often stressed on the second syllable. Unfortunately, though, the dictionary lacks data on the moving stress in the eastern and western dialects that is found in the monograph. For the southern dialects, we regularly see only one verb wordform — that of the present third person plural. This form allows us to distinguish verbs from the first or second accent paradigms; the stress placement for these paradigms is indicated in [Munkácsi 1894].

Published materials by A. Kannisto also show stress placement for southern dialects of Mansi. In Kannisto’s most recent dictionary (1982), the stress in two-syllable words and word-stems in Tavda Mansi always falls on the second syllable, for both verbs and nouns. This fact is also noted in the monograph [Honti 1975: 15].

Thus, the data reported in B. Munkácsi and A. Kannisto disagree. According to B. Munkácsi, nouns in Tavda Mansi are always stressed on the first syllable; according to A. Kannisto, the stress falls on the second syllable. For the verbs for which A. Kannisto notes fixed stress on the second syllable, B. Munkácsi lists either first- or second-syllable stress in the present third person singular form.

At present, we do not know how to resolve these controversies. It is known that A. Kannisto’s materials are phonetically more accurate than those of B. Munkácsi. For this reason, Honti’s (1982) inter-dialectal comparison of Mansi phonetics was conducted on the basis of A. Kannisto’s materials. No analogous comparison has been carried out using data from B. Munkácsi. On the other hand, as will be shown below, regular correspondences can be established between our field data, the archival records of M.A. Castrén, and Tavda materials collected by B. Munkácsi. Therefore, it seems wrong to simply dismiss the data in [Munkácsi 1894, Munkácsi, Kálmán 1986] as erroneous with respect to stress placement. We cannot exclude the possibility that the publications [Munkácsi 1894, Munkácsi, Kálmán 1986] and [Kannisto 1982] represent distinct sub-dialects of Tavda that differed, among other things, in stress placement.

The data on lexeme distribution among accent curves in Eastern and Western Mansi weren’t previously published in the scientific literature. Western Mansi is now extinct[3]; however, in 2015, while working in the National Library of Finland in Helsinki, we found a dictionary of the Upper Pelym sub-dialect (Western Mansi) [Slovtsov 1905]. This dictionary lists 424 Western Mansi lexemes, including some inflectional wordforms. For example:

Ка́йтомы ‘we-run’, *Ка́йтомъ* ‘I-run’, [Slovtsov 1905: 4];

Ка́ртэхтъ ‘they-smoke’, *Ка́ртыва* ‘we-smoke’, *Ка́ртээнъ* ‘you-sg smoke’,
Ка́ртээмъ ‘I-smoke’ [Slovtsov 1905: 15-16];

Морэ́эмъ ‘you-sg believe’, *Морэ́умъ* ‘I-believe’ [Slovtsov 1905: 5]

Кульпты́ма ‘I-leave/keep’, *Ку́льптэнъ* ‘Leave!(imv) ’ [Slovtsov 1905: 20].

An analysis of these and other examples shows that there were at least two verb paradigms in Upper Pelym: one with accent fixed on the root, and another where at least some wordforms were accented on the second syllable.

As far as we know, moving accent hasn’t previously been noted for Northern Mansi.

[3] The western dialects had disappeared already in the 1970s, according to [Rombandeeva, Kuzakova 1973].

Thus, based on the comparison of stress placement in Ob (Northern Mansi), Yukonda (Eastern Mansi) and Tavda (Southern Mansi) and Irtysh (Southern Khanty), we hypothesize that there were four accentuation paradigms at the proto-Mansi and proto–Ob-Ugric stage, which developed into three accentuation paradigms in Ob and two paradigms in Southern Khanty. The number of accentuation paradigms in the Tavda Mansi dialect is not quite clear, since the dictionary [Munkácsi B., Kálmán K. 1986] regularly lists only the third plural Present time form. Meanwhile, the data from Ob (and, looking forward into the next part of our article, Yukonda) show that stress in the third plural Present time form could be on the ending in different accentuation paradigms; a single diagnostic form does not provide sufficient information to determine the accentuation paradigm.

It is interesting that the existence of three accentuation paradigms in Ob and Yukonda is reliably supported by the data from Tavda Mansi and the southern Khanty Irtysh dialects, which have different correspondences for each paradigmatic type.

Table I

	Group I	**Group II**	**Group III.a.**	**Group III.b.**
Ob Mansi	Stress on the first syllable	Stress on the second syllable	Moving stress	Moving stress
Yukonda	Stress on the first syllable	Stress on the second syllable	Stress on the first syllable	Moving stress
Tavda [Munkácsi B.,	Stress on the first syllable	Stress on the second syllable	Stress on the second syllable	Stress on the first syllable
Khan. Irt. [Castrén,	Stress on the second syllable	Stress on the first syllable	Stress on the second syllable	Stress on the second syllable

It turns out, then, that using modern technologies to investigate the Ob-Ugric material allows us not only to record, but to analyze these data on a new level. It is clear that verifying the statements made in historical research is very important before the language material is made accessible online where detailed verification can be viewed.

Achieving this level of verification and accuracy is becoming possible now, thanks to the creation of the LingvoDoc virtual laboratory, where the functions for phonetic, morphological and etymological analysis are present. Without this resource, the results achieved on the Ob-Ugric languages would not be possible.

LingvoDoc system overview

The LingvoDoc software system provides the following features:

1. Collaborative work on dictionaries (as is provided by Google Docs or Github).

2. HTTP REST API to allow integrational functionality with any other software.

3. Web interface (reference client application) that uses REST API.

4. Flexible Access Control Lists (ACL) for collaborative editing, viewing and publishing. Each dictionary in the system can be shared to any other system user, organized for read-only, read-write and publishing purposes. Any system user without direct access to the particular dictionary can propose edits that can be reviewed by dictionary editors.

5. Personal contribution statistics.

6. Totally customizable dictionary structure. Text, sound, markup (Praat and ELAN), images, connectivity group, directed links are supported as data types.

7. Multilanguage translations for dictionaries based on the same data. All the dictionaries may contain translations to any language sharing the same media, markups, transcriptions and other data.

8. Scalable architecture (designed to utilize cloud resources for scaling).

9. Semi-offline clients with 2-way sync. A user can go to the mountains or even to another planet and still sync his data if he wants to and has an internet connection. (The user needs an internet connection for the first application launch only).

10. An ability to make **your own portals** with data that belongs to a group of users or organization; features two-way synchronization with central system capabilities.

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 use, with each dictionary hosted at one’s own institution and shared with other users or institutions at will.

13. 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.

14. Attaching dictionaries to actual **map locations**.

15. Search functionality using any plain text data in the system, with the results highlighted **on a map**.

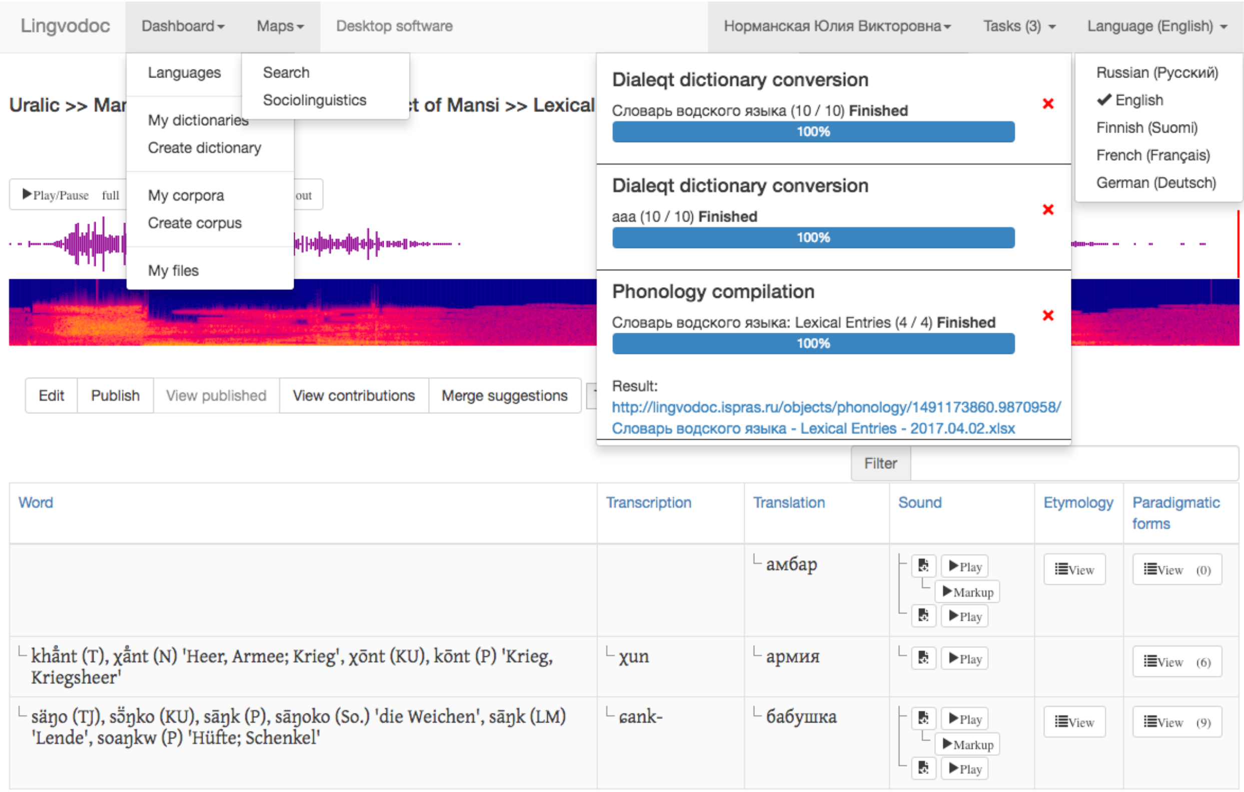
16. Duplicate search within the dictionaries, with customizable duplicate criteria.

17. Corpora in ELAN (eaf) There is a program for texts to audio markup creation named ELAN. It has its own storage format named .eaf. Our system provides the ability to store this markup files in the system and to view them inside the browser.

18. Support for **phonology computations** on the data in the system.

Sources are available for public review (the license is not permissive yet, although we are planning to migrate to Apache 2.0 as soon as we are ready; sources are currently provided for informational purposes only).

LingvoDoc system architecture



Dialeqt dictionary conversion

Словарь водского языка (10 / 10) **Finished**

100%

Dialeqt dictionary conversion

aaa (10 / 10) **Finished**

100%

Phonology compilation

Словарь водского языка: Lexical Entries (4 / 4) **Finished**

100%

Result:
http://lingvodoc.ispras.ru/objects/phonology/1491173860.9870958/
Словарь водского языка - Lexical Entries - 2017.04.02.xlsx

Word	Transcription	Translation	Sound	Etymology	Paradigmatic forms
		амбар	▶Play ▶Markup ▶Play	View	View (0)
khånt (T), χånt (N) 'Heer, Armee; Krieg', χōnt (KU), kōnt (P) 'Krieg, Kriegsheer'	χun	армия	▶Play		View (6)
säŋo (TJ), sǒ̈ŋko (KU), sāŋk (P), sāŋoko (So.) 'die Weichen', sāŋk (LM) 'Lende', soaŋkw (P) 'Hüfte; Schenkel'	ɕank-	бабушка	▶Play ▶Markup ▶Play	View	View (9)

Pic 1: Lingvodoc web interface

The main feature of the system is native support for semi-offline synchronization of user data. This feature is unique for this kind of system and is based on the concept of a composite primary key [Date 2008]. 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 key is generated.

The second basic concept is a virtual entity that doesn’t contain any data but serves as 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 key 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), an image from a manuscript, etc. Each author can have as many versions as he wants to; the system places no limits.

In data model terms, this means that we store “version” data in denormalized relational form and combine the data server-side.

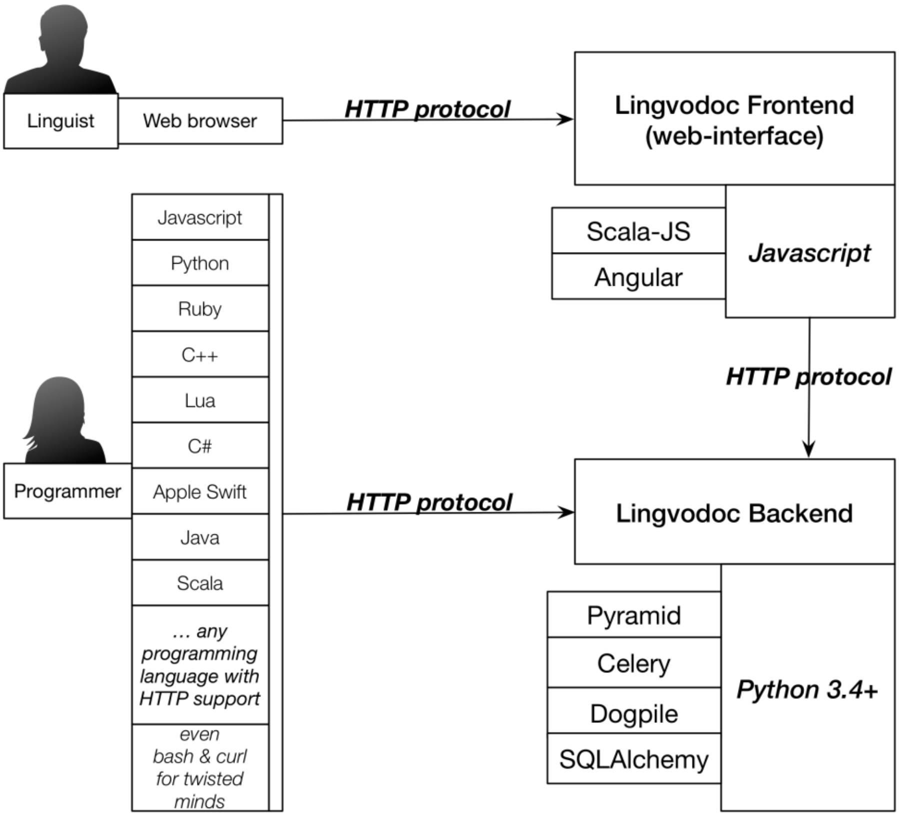
Let’s imagine, for instance, that M authors have different opinions on some object (such as the translation of a particular term). The system doesn’t limit the number of listed translations, provided each of the M authors has corresponding rights for the dictionary.

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.

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 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. All the data in the system are accessible via REST API and returned in JSON representation. The system architecture looks like this from the user’s point of view:



Pic 2: LingvoDoc architecture overview

Once you've been granted author (or owner organization) permission, you can gain any kind of access to the system.

Each author of any dictionary has a right to distribute access to his data to 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. Multilayer audio-text corpora.

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 on data 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).

LingvoDoc phonology computations

Lingvodoc supports compilation of vowel phonologies for dictionaries with sound and markup data. Given a dictionary with a set of paired sound and markup files, sound files in uncompressed WAV format and markup files in Praat TextGrid format, vowel phonology compilation proceeds as follows:

1. For each sound/markup pair, the longest vowel sound and the vowel sound with the highest intensity are determined.

2. First and second formants are computed for each distinct vowel sound found at the first step.

3. Formant data is grouped by vowel; each vowel's dataset is modelled as a bivariate normal distribution, and then a formant plot is created from the data.

Algorithms used for computation of sound intensity and formants mimic corresponding algorithms used by Praat:

- http://www.fon.hum.uva.nl/praat/manual/Sound__To_Intensity___.html
- http://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html

To compute a sound's intensity, it is first separated into chunks 0.0125 seconds long. For each chunk, an intensity value is computed by squaring the sound values, convolving them with a Gaussian window with an effective duration of 0.05 seconds (total duration of 0.1 seconds), and computing their sum. The final chunk intensity value is C * S_1 / S_2, where S_1 is the sum of the windowed squared sound values, S_2 is the sum of the window values, and C is the constant 2.5e9. Finally, the intensity of the whole sound is computed as the intensity corresponding to the mean energy averaged across the sound's chunks.

To compute the sound's formants, it is separated into chunks 0.00625 seconds long. For each chunk, the first and second formants are computed, and then the sound's first and second formant values are computed as averages of the corresponding chunk values.

Before computation of the formants, the sound is resampled to a 11025 Hz sampling frequency using NumPy's (http://www.numpy.org) FFT routines, and then filtered with a high-pass filter to emphasize frequencies higher than 50 Hz. The chunk's formant values are computed by convolving the sound segment (centered on the chunk) with a Gaussian window with a duration of 0.05 seconds (effective duration 0.025 seconds). Next, the LPC coefficients of the resulting sequence are computed using Burg's method, and finally the first and second formant values are computed as the two lowest frequencies corresponding to the roots of the LPC's characteristic polynomial.

Results of formant computation are cached for each sound/markup pair.

First and second formants of each distinct vowel sound are treated as two-dimensional vectors. After grouping computed formant vectors by vowel, the formant vectors for each vowels are modelled as a bivariate normal distribution. Mean and covariance matrix of the distributions are estimated (using a maximum likelihood estimation) as sample mean and sample covariance matrices for vowel's formant vector sets.

For each vowel, a joint formant plot constructed from the formant vector data contains a scatterplot of the vowel's formant vectors, a mean formant vector and a standard deviation ellipse, computed from the covariance matrix.

After vowel phonology compilation is completed, the computed formant dataset and constructed formant plot become available for download as a Microsoft Excel .xlsx file.

Literature.

Castrén manuscript — Manuscript of M.A. Castrén in the Finnish National Library.

Date 2008 — Date C. The relation database dictionary. Apress, 2008.

Honti 1975 — *Honti L.* System der paradigmatischen Suffixmorpheme des wogulischen Dialektes an der Tawda. Budapest – Paris. 1975.

Kannisto 1982 – Wogulische Volksdichtung gesammelt und übersetzt von *Artturi Kannisto*. VII Band. Wörterverzeichnuss zu den Bänden I-VI. Bearbeitet von Matti Liimola und Vuokko Eiras. Helsinki 1982.

Munkácsi 1894 — *Munkácsi B.* A Vogul nyelvjárások szóragozásukban ismertetve, Budapest, 1894

Munkácsi, Kálmán 1986 — Wogulisches Wörterbuch / Gesammelt von *Munkácsi B.* Geordnet, bearb. undhrsg. von *Kálmán B.* Budapest, 1986

Nikolaeva 1999 — *Nikolaeva I.* Ostyak texts in the Obdorsk dialect. Wiesbaden. 1999. (Studia uralica, Bd. 9).

Normanskaja 2014 — *Normanskaja Ju.* The system of accent in the Nizjam and south dialects of Khanty // Linguistica Uralica, 4 (50), pp. 283–302.

Normanskaja 2015a — *Normanskaja Ju.* System of moving stress in Mansi verbs and its external correspondences. Part I. Ob dialect of Mansi language // Ural-Altaic Studies, 2(17), pp.

51–66.

Normanskaja 2015b — *Normanskaja Ju.* System of moving stress in Mansi verbs and its external correlations. Part I. Yukonda dialect of Mansi // Ural-Altaic Studies, 3(18), pp. 88–103.

Rombandeeva, Kuzakova 1973 — *Rombandeeva E.I, Kuzakova E.A.* Dictionary Mansi-Russian and Russian-Mansi. Leningrad, 1982.

Sajnahova 2012 — Sajnahova A.I. Dialectology Mansi language. Khanty-Mansijsk, 2012

Slovtsov 1905 — *Slovtsov K.* The experience of the Russian-Vogul dictionary and translations into language of the Voguls ' (compiled by a priest of the Pelym Church). Tobolsk, 1905.