Acta Linguistica Academica Vol. **64** (2017) 3, 397–415
DOI: 10.1556/2062.2017.64.3.5

Corpus-oriented lexicographic database for Beserman Udmurt

Timofey Arkhangelskiy
Universität Hamburg,
Alexander von Humboldt Foundation
timarkh@gmail.com

Natalia Serdobolskaya
Russian State University for the Humanities,
Moscow State University for the Humanities
and Institute of Linguistics, RAS, Moscow
serdobolskaya@gmail.com

Maria Usacheva
Moscow State University
mashastroeva@gmail.com

Abstract: Beserman Udmurt documentation project is a long-term undertaking aimed primarily at collecting lexicographic and corpus data in the field. During our work on the project, we developed a pipeline for collecting, annotating and publishing our data. In this paper, we describe this pipeline and present the online web interface we developed for providing public access to Beserman materials. We use TLex lexicographic software for working on the dictionary and Fieldworks FLEX for annotating the corpus. After the data have been annotated, they are exported to XML and stored in the online web interface, where these two types of data become interconnected and searchable. We propose solutions to challenges that arise in projects of such kind and reflect on various constraints imposed on lexicographic databases being developed in long-term projects aimed at description of underresourced languages. We suggest that the proposed pipeline and the web interface we developed could be employed by similar projects dealing with other minority languages. The web interface based on the database and a corpus of oral Beserman texts is available online at beserman.ru.

Keywords: lexicography; Udmurt; Beserman; online dictionary; spoken corpus

1. Introduction

One of the principal parts in documenting an underresourced language or dialect is gathering lexicographic information and compiling a dictionary of the language. The success of the lexicographic work depends heavily on developing optimal structure for storing the data and using appropriate tools, i.e., lexicographic software. Of course, choosing suitable tools and

2559–8201/$ 20.00 © 2017 Akadémiai Kiadó, Budapest

data templates depends, in turn, on the data itself and on the ultimate goals of the lexicographers, e.g., who is the target audience of the dictionary, whether it will exist on paper or in digital form, etc. Our paper looks into the challenges and solutions that arise in the course of a long-term project aimed at detailed description of an endangered, but still used dialect. The presumed target audience includes researchers interested in the dialect, but also the speakers, who could use the dictionary for preserving and passing on their language. As a consequence, the database and the tools used in such a project should allow exporting the dictionary to a print-compatible format (suitable for most speakers) and to an online web interface with the possibility of search (primarily for the researchers, but also for younger heritage speakers who could use it to improve their command of the language). Since the corpus allows users and the authors of the dictionary easy access to thousands real-life usage examples, we believe that the work on dictionary should be accompanied by corpus collection, and there should be a way of interconnecting the two databases. The paper discusses the pipeline we use for dictionary and corpus collection, which involves a lexicographic database stored in TLex, a corpus database stored in Fieldworks FLEX, and an online interface developed by our team that takes XML input from these two databases and unites the data in a single web interface. We will specifically focus on our experience of dictionary development in TLex and the interconnection between the dictionary and the corpus.

The paper is based on our data and observations obtained during an ongoing fieldwork project aimed at description of the Beserman dialect of the Udmurt language (Uralic > Permic), which started in 2003, the authors of this paper being currently its key participants. Beserman is spoken by a relatively small ethnic group (according to the 2010 census, there are 2201 people identifying themselves as Beserman) who live mainly in NW Udmurtia, Russia. Beserman is usually classified as a dialect of Udmurt language which is characterized by an unusual combination of specifically Beserman language phenomena (concentrated in vocabulary and phonology) with certain traits of Northern and Southern Udmurt dialects, mostly morphological and phonetic (see Teplyashina 1970; Kel'makov 1998; Lyukina 2008). In spite of small number of speakers, the dialect remains to be the main means of every-day communication in Beserman villages, at least for the older generation; however, it is usually not passed on to the new generation nowadays. The dialect is unwritten, partly due to the fact that standard Udmurt orthography is unsuitable for it because of the differences in phonological systems. The lexicographic database we are describing is an ongoing project aiming at publication of a full-fledged paper dictionary, which could help to develop written language and encourage younger ethnic Besermans to learn and use their language. A preliminary short version of the dictionary was published recently (Kuznetsova et al. 2013), half of the dictionary containing non-verbal lemmata was published as a thesaurus (Usacheva et al. 2017), and the current versions of the dictionary and the corpus are available at beserman.ru.

The dictionary size so far is 5220 entries, which includes 1886 nouns, 2322 verbs, out of which 1011 are non-derived, 353 ideophones, 454 adjectives and adverbs, and 225 words of other categories. The fully annotated and publicly available part of the corpus comprises about 70,000 tokens, with about 40,000 more tokens transcribed but not yet glossed. The data for the dictionary was collected through elicitation and from the corpus and entered in the database by linguists. The web interface allows searching words both in the dictionary (in one of several transcriptions) and in the corpus.

2. The lexicographic database

An adequate representation of lexicographical data requires that certain portions of the data have hierarchical or tree-like, rather than plain, structure. For example, in most dictionaries there is a hierarchy of meanings, sub-meanings etc., the hierarchy of grammatical tags, etc. (Atkins & Rundell 2008; Apresyan 2009). Hierarchical organization is also supported to some extent in all widespread formats for storing lexicographic data. As an example, the TEI format allows having several homonyms (`<hom>` tag) in a dictionary entry, and each homonym, in turn, may have several senses (`<sense>` tag) (Budin et al. 2012). However, there is software that does not support such structuring or has too strict constraints on its usage, e.g., LexiquePro or the dictionary module of SIL Fieldworks FLEX.

The software we chose for our project is TLex, a tool which allows the user to create their own hierarchical templates for dictionary entries. We believe the template system it offers is, on the one hand, flexible enough to embrace most of the phenomena we could want to describe in our data, but, on the other hand, sufficiently rigid and structured in order to be easily processed and exported to XML (for the online interface) or rich text document (for printing). We also found extremely useful the possibility of collaborative work with the dictionary stored as a Postgres database on a server. We are going to present particular hierarchies and solutions used in the TLex database of the Beserman lexicographic project.

2.1. Hierarchy usage in the project

Apart from general possibility of storing the lexicographic material in a tree-like fashion, there are other examples of usage of hierarchies in our data. One particular case is conditional usage of fields. For example, in Beserman, certain nouns have a separate oblique[1] stem, cf. *š'in'* 'eye.NOM' – *š'in'm-ə* 'my eye (eye.OBL-P.1SG)', *puš* 'inner.space.NOM' – *pušk-a-z* 'into the inner space (inner.space.OBL-ILL-P.3SG)'. Therefore, the subfield "oblique stem" is only required if the field "part of speech" has the value "noun" and should be absent for other values (e.g., "verb", "adjective", etc.), which can be described in the entry template.

Another example is editing whole branches of the data tree rather than individual fields. For instance, usage examples in dictionaries are most often attached to specific meanings of lexical entries, e.g.,

(1) *iz* 'stone'
Example: *iz š'əres* 'stone road'

Bilingual dictionaries also include translations of usage examples. Hence, if a lexical entry has two meanings, 1 and 2, and each of them is illustrated by usage examples, the subfields "example 1 for the meaning 2" and "translation of the example 1 of the meaning 2" must be present only if the meaning 2 is present. If, at some point in the course of developing the dictionary, this meaning is deleted or hidden from the online view (which is relevant for the work in progress, when lexicographers at some stages need to hide non-verified information from the user) the example and its translation must be deleted (or hidden) as well. In other words, it is the whole subtree that is deleted or 'turned off', rather than a single node. This kind of operations is also supported in TLex (by contrast with, e.g., LexiquePro where the usage examples and the translations would remain untouched after the corresponding meaning has been removed).

[1] As one of the reviewers correctly pointed out, the term "oblique" that we currently use might be misleading here, since the choice of the stem is conditioned by the morphophonology (it normally appears before suffixes that start with a vowel) rather than morphology.

2.2. Flexibility of the hierarchical organization of database subfields

It is important that the hierarchical structure of the database be flexible, in terms of both the possibility of designing a complex entry structure and the possibility to alter it at any stage of the project.

During the ongoing work of the lexicographers, especially in long-term projects, hierarchical connections are regularly reviewed and reorganized. To avoid re-entering the data in such cases, the system must be flexible enough to preserve the information that has already been entered. For example, in many dictionaries the idiomatic expressions are represented as two subfields, the expression itself and its translation:

(2) *pel’* – ‘ear’
Idiom: *pel’az punə̂nə̂* – ‘memorize’ (lit. ‘put into the ear’)

However, in the Beserman project it was decided at some point to provide usage examples for the idioms, e.g.,

(3) *kət* ‘belly’, *šed’ə̂nə̂* ‘be found, be caught’
Idiom: *kət šed’ə̂nə̂* ‘become pregnant’
Example: *Solə̂ kət šed’iz* ‘She became pregnant’ (lit. ‘a belly turned up for her’)

The usage examples are needed because for some idioms the translation does not give enough information on the use of the idiom. In the example above, it is the syntactic structure of the sentence that cannot be clear if we only see the idiom: the noun phrase referring to the person who becomes pregnant is introduced by the dative case, while the noun ‘belly’ remains in the nominative.

Hence, the structure of subfields “idiom X – translation of the idiom X” has been changed to the following structure:

idiom *X* ⇒ translation of the idiom *X*
⇓
example *Y* to the idiom *X*
⇓
translation of the example *Y*

This change was made at the point when a large number of idioms had already been entered into the database. However, as the only thing that had to be changed, was allowing the Idiom node to have Example nodes as its children, the change did not affect the existing data.

The structural changes, however, can potentially be much more complex than in the aforementioned example. More radical examples of template restructuring stem from the increase of our knowledge about the documented dialect.

In Beserman, there is a class of relational nouns,[2] or inflected postpositions, which are used mostly in forms of local cases. These forms can have different distributional (part-of-speech) properties. Thus, the relational noun *vəl* 'top, surface' has:

- the illative form *vəl-e* 'to the top, to the surface (top-ILL)' which can function as a spatial postposition, non-spatial postposition or as an adverb;
- the prolative form *vəl-t'i* (top-PROL) which is a postposition with two meanings, '(moving) on the top, on the surface' or '(moving) above the top, above the surface';
- the elative form *vəl-əš'* which is a postposition with one meaning 'from the top, from the surface (top-EL)';
- the recessive form *vəl-laš'en* 'from the top (top-RECESS)' which function only as an adverb, etc.

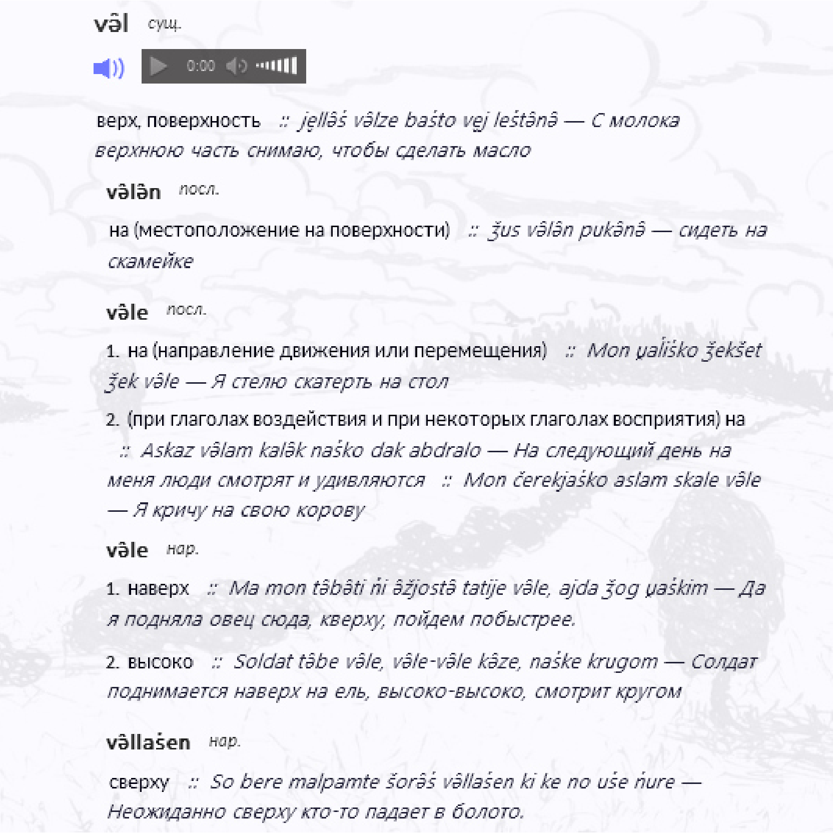
In literary Udmurt, the corresponding items are treated as postpositions; all their spatial case forms are given in Udmurt dictionaries as parts of the headword lists (see, for example, Kirillova et al. 2008). At the initial stage of our work we decided to follow the same approach. However, after several years of fieldwork we found out that in the Beserman dialect there are about 20 relational nouns most of which have forms of all or almost all local cases, and about ten postpositions which have fewer spatial forms. Since there are ten local cases in Beserman, the headword list of the dictionary has grown significantly: we had to add more than 200 entries in it. Another problem was the fact that the nominative form of a given relational noun and its local case forms are in fact used differently. The former is used quite seldom and behaves like a noun, while the latter are very frequent and behave like postpositions and/or adverbs but nevertheless retain strong

[2] Relational (or relator) nouns are defined as simple nouns (i.e., having morphological and syntactical properties of nouns) that describe spatial/temporal relations and are therefore used similarly to adpositions (cf. Starosta 1985). We present the arguments for treating these units in Beserman as relational nouns can in Arkhangelskiy & Usacheva (2015).

links with the former. It became evident that the relational noun together with its case forms should have a special status, being connected to each other more tightly than different dictionary entries, but still retaining a great deal of independence. The solution we proposed was including an additional level of sub-lemmata in our database and declaring all the local case forms of relational nouns and postpositions to be sub-lemmata of the corresponding entries (Figure 1, overleaf). The resulting XML for such entries has the following form:

```
<Lemma>

<Lemma.LemmaSign>və̄l</Lemma.LemmaSign>

...

<SubLemma>

<SubLemma.LemmaSign>və̄lə̄n</SubLemma.LemmaSign>

...

</SubLemma>

...

</Lemma>
```

Another reason for having a flexible organization of the database is the permanent change of the dialect. Lexicographic projects tend to be long-term, sometimes occupying a span of several dozen years. One important, but often overlooked, consequence of that is the fact that the databases used by the participants of such projects should have the tools allowing them to document the changes the dialect is undergoing during the period of documentation. In the 13 years of our project, we have had several cases of such change. For example, in our dictionary there is a lexical entry *vu veš'* 'small black pebble, small black shingle'. It is an obsolete word, which we recorded from only two speakers in the beginning of the project. One of the speakers has passed away since. The other is very old and has forgotten many Beserman words including the word in question. All other Beserman speakers do not know the old meaning and insist instead that the word means 'candock buds' (note that the purpose still coincides with the variant we got from the two elder speakers: the denoted object was collected and used for making beads). The option of giving the two variants as two equally possible meanings of the word (with the label "obsolete"



və̂l *сущ.*

верх, поверхность :: *jẹllə̂ś və̂lze baśto vẹj leśtə̂nə̂ — С молока верхнюю часть снимаю, чтобы сделать масло*

və̂lə̂n *посл.*

на (местоположение на поверхности) :: *ǯus və̂lə̂n pukə̂nə̂ — сидеть на скамейке*

və̂le *посл.*

1. на (направление движения или перемещения) :: *Mon ụaĺiśko ǯekšet ǯek və̂le — Я стелю скатерть на стол*
2. (при глаголах воздействия и при некоторых глаголах восприятия) на :: *Askaz və̂lam kalə̂k naśko dak abdralo — На следующий день на меня люди смотрят и удивляются* :: *Mon čerekjaśko aslam skale və̂le — Я кричу на свою корову*

və̂le *нар.*

1. наверх :: *Ma mon tə̂bə̂ti ńi ə̂ǯjostə̂ tatije və̂le, ajda ǯog ụaśkim — Да я подняла овец сюда, кверху, пойдем побыстрее.*
2. высоко :: *Soldat tə̂be və̂le, və̂le-və̂le kə̂ze, naśke krugom — Солдат поднимается наверх на ель, высоко-высоко, смотрит кругом*

və̂llaśen *нар.*

сверху :: *So bere malpamte šorə̂ś və̂llaśen ki ke no uśe ńure — Неожиданно сверху кто-то падает в болото.*

Figure 1: Example of a relational noun entry və̂l ‘top, surface’ with its sublemmata in the web interface.[3]

attached to one of them) is rejected by Beserman speakers who have seen the preliminary version of the dictionary: they dislike it when a meaning which (as the lexicographer says) has disappeared is on the same level with the actual one. The flexible structure of the dictionary entry allowed for a good compromise settlement. We introduced a new Boolean property “IsObsolete”, which can be a part of any Word sense node. We added this

[3] The metalanguage of the dictionary is Russian, since it is intended for the Beserman who are fluent in Russian. The figure shows the head word with its translation ‘top, surface’ and a number of sublemmata, each of which can have several part-of-speech blocks (e.g., *və̂le* ‘top-ILL’ can be used both as a postposition and as an adverb). All entries and subentries are accompanied by usage examples and their translations.

property with the value “True” to such obsolete senses. Structurally, therefore, these nodes are still word sense nodes. However, when exporting the dictionary for printout and for online publishing, such nodes can be treated differently: while in a speaker-oriented printed version they will probably be rendered in small font, in a separate paragraph following other senses, etc., in an online system they all word sense are represented uniformly regardless of the value of this property.

However, the freedom in designing the data structure should not be absolute. When dealing with the issues like those described above, TLex provides a reasonable equilibrium in the tradeoff between flexibility, which allows us to shape the database structure for most of the tasks which emerge during the project and modify it “on the fly”, and order, which is necessary for various tasks connected to automatic processing of the dictionary data. The flexibility is achieved by the possibility of creating infinitely customizable templates for the entries. Nevertheless, the data remains uniformly structured because every entry must conform to the template, rather than contain arbitrary fields arbitrarily related to each other, and the nodes of a template must conform to the rules for one of the predefined entity types, such as “sense”, “translation equivalent”, etc. The existence of a single template is crucial for tasks such as converting the dictionary to a printed form or using it with a searchable online interface, whereby a server side script renders entry pages using an HTML/CSS template.

2.3. A flexible system of hyperlinks

Another important feature supported by TLex is establishing labeled links (“references”) between pairs of lexical entries. For example, the verb *mə̂kə̂rtə̂nə̂* ‘bend, bow’ has references to the following lexical entries:

(4) MULTIPLICATIVE (derivation): *mə̂kə̂rjanə̂*
ITERATIVE (derivation): *mə̂kə̂rtə̂lə̂nə̂*
DETRANSITIVE (derivation): *mə̂kə̂rč’ikə̂nə̂*
SYNONYM: *n’akə̂rjanə̂* ‘bend’

There are special hyperlinks for the compounds and lexical entries that function as parts of compounds. For example, the verb *potə̂nə̂* ‘to go out’ has references to the following lexical entries as PART OF COMPOUND:

(5) *šum potə̑nə̑* ‘rejoice’,
mən’ potə̑nə̑ ‘smile’,
məl potə̑nə̑ ‘be enthusiastic to do smth.’,
vož potə̑nə̑ ‘be angry’,
žal’ potə̑nə̑ ‘pity’

One of the drawbacks of the software we are using is the impossibility of establishing links between arbitrary elements of the tree rather than between entire entries. For example, both verbs *emjanə̑* and *bə̑dtə̑nə̑* can mean ‘heal’. However, for *emjanə̑* it is the main meaning, while for *bə̑dtə̑nə̑* it is just one of the meanings (the others are ‘finish, end’ and ‘kill, do away with’). Hence, the link of the kind “synonym” would be more accurate if it connected one particular meaning of *bə̑dtə̑nə̑* to the whole lexical entry *emjanə̑* rather than two lexical entries. In some cases, it is even required to make hyperlinks to parts of subfields. Thus, in Beserman there are four synonyms denoting force. One of them is *kə̑nar* with three main meanings: ‘1. energy, (vital) force; 2. power (of machines); 3. (military) forces’, another is *kat’* ‘energy, (vital) force; health, vivacity’. The hyperlink “synonym” should be made between the first meaning of *kə̑nar* and the part of the unique meaning of *kat’* ‘energy, (vital) force ’. In TLex, such links between nodes with different status are theoretically possible, but have several significant restrictions: for example, it is impossible to add a link to a field which was added to the entry template at a later stage of its development rather than existed from the beginning. Currently, in such situations we either ignore the corresponding connection, or make it a reference to the whole lexical entry, which is inaccurate.

3. The corpus of texts and its interlinks with the dictionary

Whenever possible, dictionary entries should be illustrated with real usage examples taken from spontaneous speech, rather than with artificial ones. This can only be achieved by collecting a sufficiently large corpus of the language in question. During the fieldwork on the Beserman dialect, we have recorded and analyzed a large number of oral texts in Fieldworks FLEX.[4] This tool is a de-facto standard in language documentation projects which involve collection of small-scaled oral corpora as it provides numerous options for both morphological annotation of the data and search. Since our choice of this environment for annotating texts is rather typical for projects

[4] Apart from the authors of the paper, many other participants of the project were involved in its development.

like ours, we do not focus on describing our experience with FLEX. It must be noted however that using FLEX and manual annotation in general is only feasible for spoken corpora whose size is comparable to ours. If one chooses to integrate a dictionary with a substantially larger corpus (which is often possible in the case of standard literary languages), automatic or semi-automatic annotation of the corpus would be required.

Of course, using two unrelated pieces of software for dealing with the dictionary and the corpus has its downsides because one has to develop their own tools for interconnecting the information in them. Nevertheless, in our case this obstacle cannot be overcome since the tlCorpus, sister project of TLex, is merely a concordance software which does not provide capabilities for morphological annotation, while the dictionary module of Fieldworks FLEX was insufficiently flexible for our purposes, as was stated above. Currently, there are about 70,000 tokens in the fully glossed corpus, with around 40,000 more tokens transcribed, but not glossed. All the information on the lexical entries in the TLex database is being verified against the corpus material.

Apart from the verification of the word senses and translations, the corpus is an important source of usage examples for the dictionary. There are two possible approaches. One option is for the lexicographers to carefully select the examples manually, in order to avoid ambiguousness, multiple examples of the same word sense, etc. Alternatively, the lexicographers could list all the occurrences of every given word in the corpus to give the reader a feeling of how the word is really used in speech, including the frequency distribution of its meanings (corpus approach to lexicography, see e.g., Facchinetti 2007 and Hanks 2009). The advantages and disadvantages of both approaches are compared in 3.1 below.

3.1. Problems with adding corpora examples directly to the database

It has become a world standard for dictionaries to be based on material from electronic corpora, see Atkins & Rundell (2008, 48–50); Zeljko (2009); Cobb (2003); see also Ranchhod (2005) and Breen (2003) on the tools for extracting corpora examples for a dictionary. In this section we are going to consider the specific problems arising when using a corpus of oral texts recorded from native speakers of a minority language.

In general, corpora are regarded as the most reliable source of examples for a dictionary. This is based on the following observations. First, corpora can contain units which are difficult to study by elicitation. Corpus examples abound with discourse particles and adpositions, archaic words and idioms. These are often difficult to study by elicitation, as native speakers find it difficult to explain the meanings of the particles and provide examples of their use, and obsolete words are often unknown or used differently by different native speakers. Second, corpora examples are generally believed to be "natural", while elicited examples are influenced by the metalanguage and may sound "artificial".

However, using corpus examples can have its disadvantages, which arise both from properties of corpora in general and, specifically, from properties of oral corpora for the minority languages. First, it concerns the composition of the headword list. The most frequent words in any text, such as pronouns and other closed classes, only cover a very small part of the headword list due to the Zipf's law (Zipf 1949). All closed class items constitute less than 4% in the Beserman dictionary, while many open class words (nouns, verbs, etc.) are infrequent in texts. These classes of words, however, make the majority of the headword list (about 90% in Beserman dictionary). Those words that are infrequent or do not appear in corpus at all still make for a large portion of the headword list and have to be studied by elicitation.

Certain features characteristic for oral corpora make it even more problematic to use corpus examples in a dictionary. Specifically, the following general properties of oral speech (Kibrik & Podlesskaya 2009) might generate obstacles for inclusion of the corpus examples:

- incomplete sentences, autocorrection
- abundance of hesitation markers and placeholders
- null realization of previously mentioned participants
- abundance of anaphoric references
- reduction of sentences on the base of common knowledge of the interlocutors and of a broad context

Consider the following example from the Beserman corpus:

(6)	Və̂li-ja-z	pesok-o=no	mil'am,	pesok-a-z	ez=no	pot-ə̂,
	upper-LOC-P.3SG	sand-ATTR	we.GEN	sand-LOC-P.3SG	NEG.PST=ADD	come.out-SG

mar-ke	tue	**marə̂m,**	mijə̂m	**marə̂m**	**kar-i-z...**
what-INDEF	this.year	HES	last.year	HES	do-PST-3SG

kwaš'm-i-z	leš'a	soku	pəš'-en.
dry.up-PST-3SG	it.seems	then	hot-INS

lit. 'On the top we have sandy, and in the sand it did not come out, somehow this

year, er, last year it, er, dried up it seems, by that heat.'
'On top we have sandy [soil], that's why [potatoes] did not grow this year in the sand, and last year all of it dried up because of that heat.'

This sentence contains incomplete clauses, repetitions, two hesitation markers, and two null referents ('soil' and 'potatoes'). All of these make it a long and complicated example (requiring many explanations and comments) when taken out of the larger context.

The fact that the interlocutors make extensive use of common knowledge leads to the necessity of long comments. Consider the following example:

(7) Man'et-ez=ke=pe ug mə̂n-ə̂, berlan'=pe berek-č'ik-o-z.
coin-P.3SG=if=CIT NEG go-SG back=CIT return-DETR-FUT-3SG
'If the coin is not going, [the soldier] will come back, they say.'

The sentence is taken from the text about the traditional rituals that used to be performed when sending a conscript to the army. A piece of fabric had to be hammered into the beam under the roof. The fabric was hammered with a coin on top of it. If the coin did not stay in the beam ("did not go"), it was taken as a sign that the conscript would return home safely. In order to explain this example properly, a large comment is needed.

The fact that the interlocutors use broad context leads to the necessity of providing long context when citing corpus examples. Consider the following sentence:

(8) Pal'l'an pal-iš'en, ben=a?
left side-EGR yes=Q
'From the left side, isn't it?'

In order to interpret this example correctly, the reader should have access to at least several preceding sentences: "Well, there are two windows there. – Yes, to which of them should I put that one [the card with a picture of a cow on it]? – [To the] one of them is above the door, higher than the door. – From the left side, isn't it?"

Certain problems also arise when interpreting examples from the corpora. It has been shown by Kibrik et al. (2010) that even native speakers exhibit high level of divergence when they analyze the corpora examples (the study is based on the experience of referential interpretation tagging of Russian texts by native speakers). A variety of possible interpretations is also observed in the analysis of lexical phenomena. First, not all the contexts allow to differentiate between two meanings of one and the same lexical entry. In the following sentence that contains the verb *vəš'aš'kə̂nə̂* 'to cross oneself, to pray'[5] it is impossible to say for sure which of the two senses was meant by the speaker:

(9) Pop **vəš'aš'k-ə̂l-i-z**, mol'itva lə̂ǯ'-ə̂l-i-z.
priest cross/pray-ITER-PST-3SG prayer read-ITER-PST-3SG
{A priest used to come in the beginning of the Korban feast.} 'The priest would **cross himself/pray**, would recite a prayer.'

Probably, the discussed character both crossed himself and prayed, but the exact interpretation and translation is hindered by the use of a polysemic word. We have to ask the author of the text to make sure which one they had in mind. Such a context is, hence, not a good diagnostics to show the peculiarities of the word's use. A similar problem is observed in the following example containing the noun *č'ə̂rtə̂* 'neck, throat':

(10) **Č'ə̂rtə̂-ze** so paš' pot-t-em və̂lem=n'i.
neck/throat-ACC.P.3SG that hole come.out-CAUS-PST2 be.PST2=already
{The wolf bit the sheep.} 'It made a hole in its **neck/throat**.'

In this context, both of the interpretations are valid (as the sheep had a hole in both its neck and its throat). It is hence unclear, what meaning of this lexical entry it is illustrating.

The examples above address another problem with the corpora examples. When looking for examples for the dictionary, a lexicographer often endeavours to find the more illustrative ones, that is the ones that show exactly the meaning in question (neck or throat, to lock or to close). If s/he inserts the ambiguous examples as given above, it is going to confuse the reader of the dictionary. Hence, such examples are preferably omitted while compiling a dictionary database.

As we see, "blind" export of corpus examples into the dictionary of a minority language appears to be a complicated and inefficient procedure. It is more efficient to combine elicited examples with examples from the corpus, selected manually. This seems to be an optimal solution, especially for oral corpora of minority languages. Therefore, when compiling an online digital dictionary, it is not necessary to keep the dictionary database and the corpus interconnected. In the online interface we developed, usage examples entered by the lexicographers are shown separately from those

[5] Although the Besermans we work with are generally only nominally Orthodox Christian, they are familiar with both Christian and Muslim rites and rituals and distinguish between these two notions.

found in the corpus automatically, so that the user could use both kinds of examples differently while having the advantage of consulting all of them at the same page.

3.2. Export of the concordance examples into the database

The export of the concordance examples into the lexicon is easily done in FLEX, and the concordance is made automatically for all the lexicon entries. Note, however, that the interlinks between the lexicon and a FLEX text corpus are not supported in the TLex software (as well as in other tools). Therefore, after the export of the textual examples, the lexicographers decide which of them deserve to be included in the lexical database and transfer them manually, which requires lots of manual work. Manual selection of real usage examples is indispensable for a printed dictionary, both because of space constraints and the target audience. One problem of this approach is that once compiled, the set of examples does not take into account further additions to the corpus, which is being continually updated and may provide more suitable examples in a while.

3.3. The web interface for the dictionary and the corpus

The possibility of direct search of all occurrences of a dictionary entry in the corpus is more important for the researchers and is, of course, confined to the online version of the dictionary. Since the dictionary and the corpus are stored in two incompatible systems, there was no ready solution for joining their data in one searchable web interface. However, both systems allow export to XML files, which made the task of developing such an interface relatively easy, provided the lexical information is encoded uniformly in the dictionary and in the corpus (transliteration system, grammatical tags, etc.).

In 2015, we developed a web interface that integrates the dictionary and the corpus, giving the user the possibility of consulting corpus examples when looking at a dictionary entry. The solution consists of a Python (flask) backend which accepts queries from the HTML/CSS/JavaScript frontend. The XML trees of the dictionary and the corpus are parsed by the preprocessing scripts and are stored as an inverse index that for each stem indicates in which corpus sentences it could be found. Search queries available for the user include looking for a specific dictionary entry, looking for all entries whose translation contain certain Russian words, and looking for all entries that contain certain pattern (using regular expressions).

When the user inputs the query, it is relayed to the server by a JavaScript code. The server looks up the dictionary entry or entries, as well as their occurrences in the corpus, and transforms them into an HTML page using templates, which is then sent back to the user. The template does not include certain fields intended for lexicographers' use only, and the interface has a set of options allowing the lexicographers to select which items can be displayed (e.g., those which have the "Ready" status in the database). This is the point which requires that all dictionary entries conform to a single template. Also, developing such a system imposes certain restrictions on the flexibility of the system, as from now on most changes to the database structure must be paralleled in the scripts transforming the data to HTML. As the size of the corpus does not and probably will not exceed several hundred thousand tokens, which is usually the case with small non-written languages, we did not employ sophisticated corpus platforms or search systems.

By default, each dictionary entry is accompanied by up to ten randomly selected usage examples from the corpus, which are shown alongside the examples which are part of the entry. This potentially leads to a problem: the examples already present in the entry might have been taken manually from the corpus and thus may coincide with those taken from the corpus online. It is not that easy to find all such cases though, since corpus examples often undergo small cosmetic changes before being included in the dictionary. In order to minimize possible coincidences, the server compares normalized examples from dictionary entries to those found in the corpus and discards the latter if the Levenshtein distance between the two is too small.

The fact that the web interface is designed for both international researchers and native speakers means that we have to include the possibility of using multiple writing systems. While it is common for the researchers to use Latin-based script (e.g., Uralic phonetic alphabet), the native speakers use a Cyrillic-based orthography (based on Standard Udmurt orthography). The web interface allows the user to choose from multiple writing systems, provided there is a way of automatically transform queries and output between these systems. In the Beserman web dictionary, three systems are available: Uralic Latin-based; another Latin-based system used in our project, and Cyrillic-based system for the native speakers. The conversion between the three is carried out by Python functions which are called whenever the user chooses writing system different from that used in the dictionary database.

One of the proposed developments of an online search system in our case is connecting it to the Corpus of Standard Udmurt[6] so that the users could look at the usage of the literary cognates for Beserman words. Such a development will include alignment of the Beserman and literary Udmurt lemmata (Miller 2017).

4. Conclusion

Compiling a dictionary for an underresourced dialect with the intention of publishing it both on paper and digitally, imposes certain constraints on the organization of the lexicographic database. On the basis of our experience in the Beserman lexicographic project, we have demonstrated that the TLex software offers appropriate solutions in most cases. However, in large long-term projects with any lexicographic system there will be need for more capabilities than an out-of-the box system can provide, first of all when preparing an online version of the dictionary interconnected with the corpus. As we have shown, such problems can be overcome by customization of the entry templates or developing additional modules working with the XML representation of the database. The solution we developed uses XML output from TLex and Fieldworks FLEX and incorporates them in a web interface where the user can see the corpus examples when searching the dictionary. The interface is mostly language-independent and could be used in other documentation projects with similar pipelines.

[6] http://web-corpora.net/UdmurtCorpus/search

References

Apresyan, Yuriy D. (ed.). 2009. Issledovaniya po semantike i leksikografii. Vol. I: Paradigmatika [Studies of semantics and lexicography. Vol. 1: Paradigmatics]. Moscow: Yazyki slavyanskikh kul'tur.

Arkhangelskiy, Timofey and Maria Usacheva. 2015. Syntactic and morphosyntactic properties of postpositional phrases in Beserman Udmurt as part-of-speech criteria. SKY Journal of Linguistics 28. 103–137.

Atkins, B. T. Sue and Michael Rundell. 2008. The Oxford guide to practical lexicography. Oxford: Oxford University Press.

Breen, Jim W. 2003. Word usage examples in an electronic dictionary. Manuscript. Papillon (Multi-lingual Dictionary) Project Workshop, Sapporo, July 2003.

Budin, Gerhard, Stefan Majewski and Karlheinz Mörth. 2012. Creating lexical resources in TEI P5. Journal of the Text Encoding Initiative 3.

Cobb, Tom. 2003. Do corpus-based electronic dictionaries replace concordancers? In B. Morrison, C. Green and G. Motteram (eds.) Directions in CALL: Experience, experiments. Hong Kong: Polytechnic University. 179–206.

Facchinetti, Roberta. 2007. Theoretical description and practical applications of linguistic corpora. Verona: QuiEdit.

Hanks, Patrick. 2009. The impact of corpora on dictionaries. In P. Baker (ed.) Contemporary corpus linguistics. London: Continuum. 214–236.

Kel'makov, Valej K. 1998. Kratkiy kurs udmurtskoy dialektologii. Vvedenie. Fonetika. Morfologiya. Dialektnye teksty. Bibliografiya [A brief sketch of Udmurt dialectology. Introduction. Phonetics. Morphology. Texts in Udmurt dialects. References]. Izhevsk.

Kibrik, Andrey A., Dobrov Grigoriy B., Zalmanov Dmitriy A., Linnik Anastasia S. and Lukashevich Natalia V. 2010. Referentsial'nyj vybor kak mnogofaktornyy veroyatnostnyy protsess [Referential choice as a multifactorial probabilistic process]. Kompyuternaya lingvistika i intellektual'nye tekhnologi 9. 173–181.

Kibrik, Andrey A. and Vera I. Podlesskaya. 2009. "Rasskazy o snovideniyakh": korpusnoe issledovanie ustnogo russkogo diskursa ["Dream stories": A corpus study of Russian oral discourse]. Moscow: Yazyki slavyanskikh kul'tur.

Kirillova, Lyudmila E. et al. (eds.). 2008. Udmurtsko–russkiy slovar' [Udmurt–Russian dictionary]. Izhevsk: UUIYaL UrO RAN.

Kuznetsova, Ariadna I. et al. 2013. Slovar' besermyanskogo dialekta udmurtskogo yazyka [Dictionary of the Beserman dialect of Udmurt]. Moscow: Tezaurus.

Lyukina, Nadezhda M. 2008. Osobennosti yazyka balezinskikh i yukamenskikh besermyan (sravnitel'naya kharakteristika) [The peculiarities of the language of Balezino and Yukamenskoe Besermans (a comparison)]. Doctoral dissertation. Izhevsk.

Miller, Evgeniya O. 2017. Avtomaticheskoe vyravnivanie slovarey literaturnogo udmurtskogo yazyka i besermyanskogo dialekta [Automatic alignment of Literary and Beserman Udmurt dictionaries]. In Proceedings of Elektronnaya pis'mennost' narodov Rossiyskoy Federatsii: Opyt, problemy i perspektivy [Electronic literacy of the peoples of the Russian Federation: Experience, challenges and perspectives]. Syktyvkar. 109–111.

Ranchhod, Elisabete Marques. 2005. Using corpora to increase Portuguese MWU dictionaries: Tagging MWU in a Portuguese corpus. In Proceedings from the Corpus Linguistics Conference Series. Birmingham: University of Birmingham.

Starosta, Stan. 1985. Relator nouns as a source of case inflection. In V. Z. Acson and R. L. Leed (eds.) For Gordon H. Fairbanks. Honolulu: University Press of Hawaii. 111–133.

Teplyashina, Tamara I. 1970. Yazyk besermyan [The language of the Besermans]. Moscow: Nauka.

Usacheva, Maria N. et al. 2017. Tezaurus besermyanskogo narechiya: Imena i sluzhebnye chasti rechi (govor derevni Shamardan) [Thesaurus of the Beserman dialect: Nouns and auxiliary parts of speech (Shamardan village variety)]. Moscow: Izdatel'skie resheniya.

Zeljko, Miran. 2009. Improvements of dictionaries – Suggestions by Evroterm. In H. Stančić, S. Seljan, D. Bawden, J. Lasić-Lazić and A. Slavić (eds.) Future2009: Digital resources and knowledge sharing. Zagreb: University of Zagreb. 269–279.

Zipf, George Kingsley. 1949. Human behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley Press.