СОЗДАНИЕ ЛЕКСИКОГРАФИЧЕСКОЙ БАЗЫ ДАННЫХ ДЛЯ ЭЛЕКТРОННОГО СЛОВАРЯ

Архангельский Т. А. (tarkhangelskiy@hse.ru)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Идрисов Р. И. (idrisov.ru@gmail.com)
Институт языкознания РАН, Москва, Россия

Сердобольская Н. В. (serdobolskaya@gmail.com)
РГГУ и Институт языкознания РАН, Москва, Россия

Усачёва М. Н. (mashastroeva@gmail.com)
МГУ им. Ломоносова, Москва, Россия

Ключевые слова: лексикография, удмуртский язык, бесермянский диалект, онлайн-словари, лингвистические базы данных

DESIGNING A LEXICOGRAPHIC DATABASE FOR AN ONLINE DICTIONARY[1]

Arkhangelskiy T. A. (tarkhangelskiy@hse.ru)
National Research University Higher School of Economics, Moscow, Russia

Idrisov R. I. (idrisov.ru@gmail.com)
The Institute of Linguistics, RAS, Moscow, Russia

Serdobolskaya N. V. (serdobolskaya@gmail.com)
Russian State University for the Humanities, Moscow State University for the Humanities and Institute of Linguistics, RAS, Moscow, Russia

Usacheva M. N. (mashastroeva@gmail.com)
Moscow State University, Moscow, Russia

[1] The work is supported by the RFBR grant № 13-06-00179.

In the paper, we discuss various constraints imposed on lexicographic databases being developed in long-term projects aimed at description of an underresourced language and publishing the dictionary of the language both in printed form and as an online resource. We present the dictionary database for the Beserman dialect of Udmurt, collected during an ongoing fieldwork project started in 2003. The paper describes the challenges that arise in projects of such kind and the proposed solutions. For the dictionary database, we used the TshwaneLex software which provides a reasonable balance between flexibility and customizability (needed for dealing with the challenges), on the one hand, and orderedness of the data (needed for seamless interaction with the online interface), on the other. A preliminary print version of the dictionary based on the presented database was published in 2013 (Kuznetsova et al. 2013). An online searchable web interface based on the database and a corpus of oral Beserman texts, is currently under development and is expected to become available for public access by May 2015.

Keywords: lexicography, Udmurt, Beserman, online dictionary, linguistic database

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 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 in 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 widely 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). The paper discusses features of the database relevant in such a setting and describes our experience with using a software called TscwaneLex in building a dictionary of such kind.

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 > Permian), which started in 2003. Beserman is spoken by a relatively small ethnic group (according to the 2010 census, there are 2201 people identifying themselves as Beserman) occupying the basin of the Cheptsa river and the Kirov region of the Russian Federation. In the scientific literature the Beserman is considered to be a dialect of Udmurt language which is characterized by an unusual combination of specifically Beserman language phenomena (concentrated in vocabulary and phonetics) with certain traits of Northern and Southern Udmurt dialects, mostly

morphological and phonetical (see Teplyashina 1970, Kelmakov 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.

2. Organization of the lexicographical 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 marks and other (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 Language Explorer. The software we chose for our project is TshwaneLex, a tool which allows the user to create their own hierarchical templates for dictionary entries. We are going to present particular hierarchies and solutions used in the TshwaneLex 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 stem, cf. *š'in'* 'eye (NOM)'—*š'in'm-ə* 'my eye (eye.OBL-POSS.1SG)', *puš* 'inner space (NOM)'—*pušk-a-z* 'into the inner space (inner.space.OBL-ILL-POSS.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'ə'res* 'stone road'

Bilingual dictionaries also include translations of the 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 TshwaneLex (by contrast with, e. g. LexiquePro where the usage examples and the translations would remain untouched after the corresponding meaning has been removed).

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 of 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:

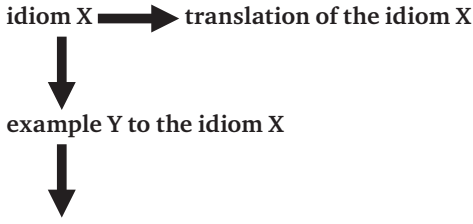
(2) *pel'*—'ear'
Idiom: *pel'az punə'nə'*—'memorize' (lit. 'put into the ear')

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

(3) *kət* 'belly', *šed'ə'nə'* 'be found, be caught'
Idiom: *kət šed'ə'nə'* 'become pregnant'
Example: *Solə' kət šed 'iz* 'She became pregnant' (lit. 'to-her a belly was found')

The usage examples are needed because for some idioms the translation does not give enough information on the use of the idiom. In the given example 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:



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.

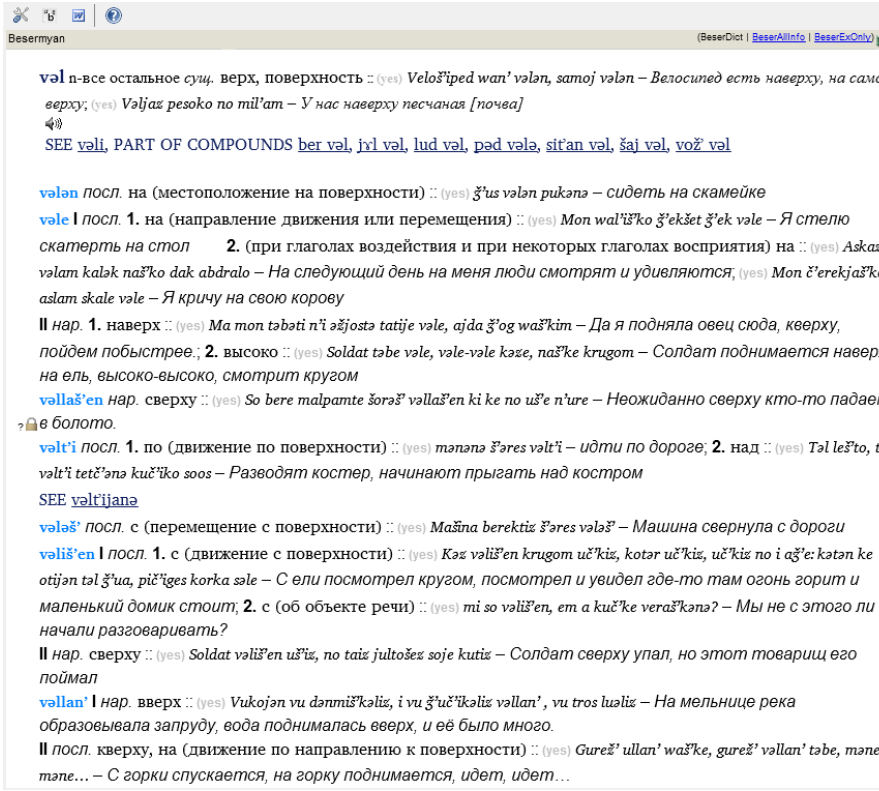
In Beserman, there is a class of relational nouns[2] which are used mostly in forms of local cases. These forms can have different part-of-speech properties. Thus, the relational noun *vəʼl* 'top, surface' has:

- the illative form *vəʼl-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 *vəʼl-tʼi*(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 *vəʼl-əʼšʼ*which is a postposition with one meaning 'from the top, from the surface (top-EL)';
- the recessive form *vəʼl-lašʼen* '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 10 postpositions which have less spatial forms. Since there are 10 local cases in Beserman, the headword list of the dictionary has grown significantly: we had to add around 250 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 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 (fig. 1). The resulting XML for such entries has the following form:

[2] Relational nouns denote the first term of relation and have a valence which must be filled at the syntactic level (Shmelev 1998: 170–171).

<Lemma>
<Lemma.LemmaSign>vəl</Lemma.LemmaSign>
...
<SubLemma>
<SubLemma.LemmaSign>vələn</SubLemma.LemmaSign>



...
</SubLemma>
...
</Lemma>

Fig. 1. Example of a relational noun entry vəl ‘top, surface’ with sublemmata[3]

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

[3] The transcription system in the dictionary currently differs slightly from the one used in the paper and will be revised before publishing the dictionary online.

during the period of documentation. In the 12 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 'small black pebble, small black shingle', however, they are sure that the word means 'can-dock buds' (note that the purpose still coincides with the variant we got from the two elder informants: 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" attached to one of them) is rejected by Beserman speakers we have shown the dictionary to: they dislike it when a meaning which (as the lexicographer says) has disappeared is written together with the actual one. The flexible structure of the word article allows for a good compromise settlement. For such obsolete senses, a separate node type could be declared. Structurally, it would behave just like a word sense (e. g. it will be a sibling of other word senses in the tree). However, when exporting the dictionary for printout and for online publishing, such nodes could be treated differently: while in a speaker-oriented printed version they can be rendered in small font, in a separate paragraph following other senses, etc., in an online system it can be represented as an ordinary word sense.

However, the freedom in designing the data structure should not be absolute. When dealing with the issues like those described above, TshwaneLex provides a reasonable equilibrium in the tradeoff between flexibility, which allows us to shape the database structure for all 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 ordered 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.

3. A flexible system of hyperlinks

Another important feature supported by TshwaneLex is establishing labeled links ("references") between pairs of lexical entries. For example, the verb *mə'kə'rtə'nə'* 'bend, bow' has references to the following lexical entries:

(4) multiplicative (derivation): *mə'kə'rjanə'*
iterative (derivation): *mə'kə'rtə'lə'nə'*
detransitive (derivation): *mə'kə'rč'ikə'nə'*
synonym: *n'akə'rjanə'* 'bend'

However, not all types of links are currently fully supported by this software. For example, both verbs *emjanə'* and *bə'dtə'nə'* can mean 'heal'. However, for *emjanə'* it is the main meaning, while for *bə'dtə'nə'* 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 *bə'dtə'nə'* to the whole lexical entry *emjanə'* rather than two lexical entries. In TshwaneLex, 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.

4. Interlinks between the database and the text corpus

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 Language Explorer[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. 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 TshwaneLex, is merely a concordance software which does not provide capabilities for morphological annotation, while the dictionary module of Fieldworks Language Explorer was insufficiently flexible for our purposes, as was stated above. Currently, there are about 75,000 tokens in the corpus[5]. All the information on the lexical entries in the TshwaneLex database has been verified according to 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, 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).

Both approaches are used in different parts of our project.

---

[4] Apart from the authors of the paper, many other participants of the project were involved in its development. Final checking and glossing was carried out by Olga Biryuk.

[5] An older version of the corpus (approx. 7,000 sentences) is available online at http://corpora.iling-ran.ru/index.php.

4.1. Export of the concordance examples into the database

The export of the concordance examples into the lexicon is easily done in Fieldworks, and the concordance is made automatically for all the lexicon entries. Note, however, that the interlinks between the lexicon and a Fieldworks text corpus are not supported in the TshwaneLex software (as well as in other tools, e.g. LexiquePro). 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.

4.2. Search in the corpus from the lexical database

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 (TshwaneLex and Fieldworks), there is no ready solution for joining their data in one searchable web interface. However, both systems allow export to an XML file, which makes 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.). The solution we are currently developing[6] includes a Python server side script which accepts queries from the HTML/CSS/JavaScript interface. The script parses the XML trees of the dictionary and the corpus, finds the item or items being queried and their occurrences in the corpus and transforms them into HTML, which is then sent back to the user. The transformation does not include certain fields intended for lexicographers' use only, and 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, it is unnecessary to employ sophisticated corpus platforms or search systems.

Of course, having such a system allows more complex queries (e. g. using regular expressions) at almost no additional cost. However, its search possibilities will still be limited to those implementable without recur to advanced corpus platforms, which would mean much additional work for achieving a goal peripheral for the project.

---

[6] A much older, preliminary version of the dictionary developed in another framework is available at http://languedoc.philol.msu.ru/~beserman/.

One of the proposed developments of an online search system in our case is connecting it to the Corpus of Standard Udmurt[7], so that the users could look at the usage of the literary equivalents for Beserman words. However, such a development will include alignment of the Beserman and literary Udmurt lemmata, which must be done, at least in part, manually.

5. 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 TshwaneLex 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 enty templates or developing additional modules working with the XML representation of the database.

References

1. *Apresyan Yu. D. (ed.),* (2009), Issledovaniya po semantike i leksikografii, Vol. I: Paradigmatika [Studies of semantics and lexicography. Vol. I: Paradigmatics], Jazyki slavyanskikh kultur, Moscow.
2. *Atkins B. T. S., Rundell M.* (2008), The Oxford Guide to Practical Lexicography, Oxford University Press, New York.
3. *Budin, G., Majewski, S., and Mörth, K.* (2012). Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries. Journal of the Text Encoding Initiative, (3).
4. *Facchinetti R.* (2007), Theoretical Description and Practical Applications of Linguistic Corpora, QuiEdit, Verona.
5. *Granger S., Paquot M. (eds.)* (2012), Electronic Lexicography, Oxford University Press, Oxford.
6. *Hanks P.* (2009), The impact of corpora on dictionaries, in P. Baker (ed.) (2009), Contemporary Corpus Linguistics, Continuum, London,pp. 214–236.
7. *Kelmakov V. 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.

[7] The Udmurt corpus was developed by Timofey Arkhangelskiy and Maria Medvedeva in 2014 and is available at http://web-corpora.net/UdmurtCorpus/search.

8. *Kirillova L. E.* (chief ed.) et al. (2008), Udmurtsko-russkij slovar [Udmurt-Russian Dictionary], UUIYaL UrO RAN, Izhevsk.
9. *Kuznetsova, Ariadna I. et al., eds.* (2013). Slovarʾ besermyanskogo dialekta udmurtskogo yazyka [Dictionary of the Beserman dialect of Udmurt]. Moscow: Tezaurus.
10. *Lyukina N. M.* (2008),Osobennosti yazyka balezinskikh i yukamenskikh besermyan (sravnitel'naya kharakteristika) [The peculiarities of the language of Balezino and Yukamenskoe Besermans (a comparison)], Diss. Izhevsk.
11. *Shmelev A. D.* (1998) Tipy "nevyrazhennykh" valentnostey[The types of "ulterior" valences], in Semiotika i informatika. Sbornik nauchnykh statey, vol. 36, Jazyki russkoj kultury, Moscow, pp. 167–176.
12. *Sinclair J. (ed.)* (1987), Collins COBUILD English Language Dictionary, HarperCollins Publishers, London.
13. *Teplyashina T. I.* (1970), Yazyk besermyan [The language of the Besermans], Nauka, Moscow.