Acta Linguistica Academica Vol. **64** (2017) 3, 383–396
DOI: 10.1556/2062.2017.64.3.4

Corpus and lexicon databases of Khanty and Mansi dialects

Axel Wisiorek
Ludwig Maximilian University of Munich
Institute for General Linguistics and Language Typology, IT Group for the Humanities
axel.wisiorek@lmu.de

Zsófia Schön
Ludwig Maximilian University of Munich
Institute for Finno-Ugric Studies
zsofia.schoen@gmail.com

Abstract: In this paper we describe the data processing procedures and the preliminary results of the project Ob-Ugric database (OUDB), a web-based framework which aims at developing corpus-based descriptive resources of Khanty and Mansi dialects. Using established language documentation and annotation tools, OUDB provides interlinked corpus and lexicon data from digitized texts as well as recent fieldwork studies in an uniform IPA-transcription together with the corresponding audio recordings thus making these less described languages of the Ob-Ugric branch of the Finno-Ugric language family accessible for researchers as well as the language community and archiving the raw data for documentation, linguistic evaluation and possible future use in building resources for language technology applications.

Keywords: Khanty; Mansi; annotated corpora; corpus-based lexical database; language documentation

1. Introduction

The need for well annotated data of endangered languages, among them the dialects of Khanty and Mansi, is evident and regarding the rapid dwindling of these languages, requirements imposed on corpora such as representativeness, genre diversity or minium size are negligible in view of the need for documentation and data availability for linguistic research (cf. Gries 2009, 1237–1238). The project *Ob-Ugric database: analysed text corpora and dictionaries for less described Ob-Ugric dialects*[1] (OUDB; July

[1] http://www.oudb.gwi.uni-muenchen.de

2559–8201/$ 20.00

2014–June 2017, Munich/Vienna) and its corpus database try to cover as much material as possible from these dialects belonging to the Ob-Ugric branch of the Finno-Ugric language family, and to enrich these language data with multiple annotational layers (phonetic, phonotactic, morphological, syntactic and pragmatic) and consequently serve as a multipurpose corpus data resource.

To achieve these objectives, OUDB primarily uses the established language documentation and annotation tools FLEx and ELAN, i.e., takes a semi-manual annotation approach as its basis. For such less described languages with small available corpora, high dialectal variability and heterogenous or non-existing orthographical standards, it is a reasonable approach to analyze a core corpus (semi-)manually as basic usage-based description for documentation purposes.[2]

Currently, the size of this semi-automatically morphological-tagged corpus is about 40,000 word tokens, with the total corpus having over 200,000 tokens in approximately 430 texts.

The fundamental database structure, the data processing routines and the PHP-based web framework including a backend for cooperating researchers were initially set up in the course of the project *Ob-Ugric languages: conceptual structures, lexicon, constructions, categories* (OUL, August 2009–July 2012), which dealt with already published written material from two Khanty (Kazym and Surgut) and two Mansi (Northern and Southern) dialects. In this initial project of the universities of Munich, Vienna, Szeged and Helsinki, the documentation and annotation software FieldWorks Language Explorer (FLEx)[3] was chosen for the task of analyzing the corpus data, including segmentation and morphological tagging. In the course of the manual tagging process, this widely used language documentation toolkit builds up a stem and affix lexicon, and the build-in

[2] These annotated data subsequently can be used to train machine learning based systems and thus obtain probabilistic, usage-based language models, which can be utilized for extending the annotated corpus or in other fields of application of language technology. In a preliminary test, we trained the decision tree based part-of-speech tagger *TreeTagger* (by Helmut Schmid, http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger, see Schmid 1994) on the Northern Mansi sub-corpus (about 14,000 tokens) which was divided in a training set of 86% (12.000 tokens) and a test set of 14%; this statistical POS-model achieves an accuracy (for words and punctuation; with a broad tagset) of 78%; for single part-of-speech-categories, it reaches better results, e.g., 83% accuracy for verbs, with a precision of 90%. With increased corpus size (based on these results and manual correction), the accuracy will increase as well (cf. Schmid 1994).

[3] http://software.sil.org/fieldworks/

morphological parser assists the researcher with suggestions based on this gradually extended lexicon (see Black & Simons 2006).

A graphic unification for each of the dialects was elaborated using IPA-characters, where the different, idiomatic writing systems of the used sources (e.g., phonetic or phonologic, latin based or Cyrilic transcriptions etc.) were brought to a common ground in the corpora, and all material used in OUL was transliterated using a unified phonological system. The need for a standardized transcription was prioritized to the original form of the texts, on the one hand for easy data handling, and on the other for accessibility to linguists outside the field of Finno-Ugric Studies.

As the number of dialects covered grew with OUDB – a cooperation between the universities of Munich and Vienna – data not only increased in volume, but also became more and more heterogeneous: while the extinct Pelym and North-Vagilsk dialects of Mansi are represented only by text editions from the end of the 19th century, the Yugan dialect of Khanty relies on sound recordings from fieldwork in the 21st century, which were transcribed using the annotation tool ELAN Linguistic Annotator.[4]

On the basis of the already established data processing workflow, the relational data model and the web framework, OUDB continued to develop these corpus and lexicon tools, with expanded filter and search possibilities, an updated interface, and enriched audio data. It features elaborated interlinear glosses of complete texts, an innovative concordancer which makes the annotated corpus data highly searchable for various patterns, as well as a corpus-based electronic dictionary, its entries directly connected with the text corpus via the concordance module. Main advantages of using a web-based service for a research platform such as OUDB are platform independency, long-term availability and easy international collaboration through the client-server model (cf. McEnery & Hardie 2011, 45ff; Hardie 2012).

The gained multipurpose language data (including the audio recordings and metadata) will be made available for download at the end of the project in an ELAN-XML-format via *The Language Archive*/MPI.

[4] https://tla.mpi.nl/tools/tla-tools/elan

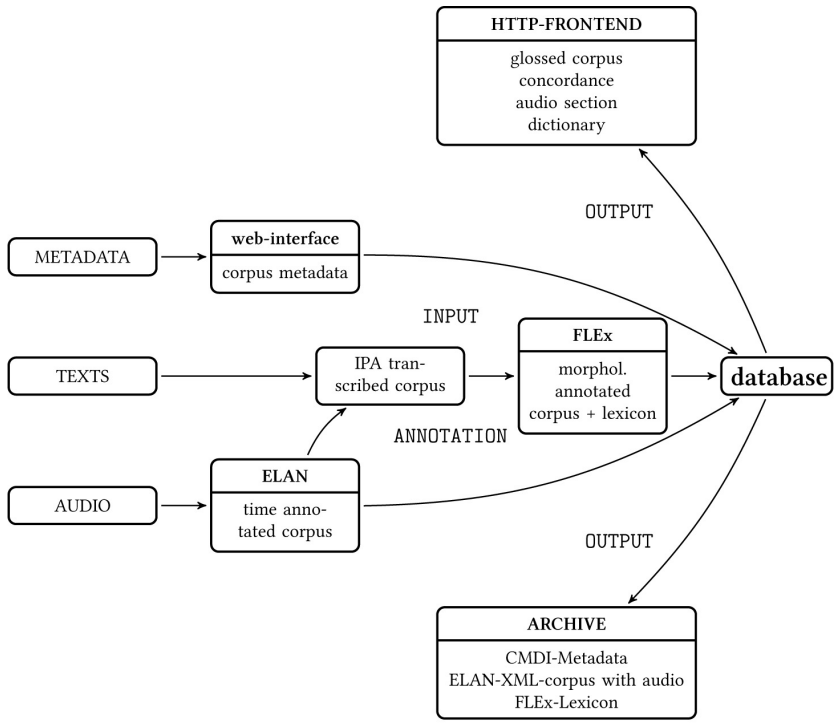

Figure 1: Data processing

2. Data processing

2.1. Data input and annotation

The following section introduces the data processing workflow and the annotation procedure, as it is illustrated in Figure 1.

The primary data is either converted to the OUDB-IPA transcription standard or (if fieldwork-audio) analyzed with ELAN (in an ASCII-transcription, which is likewise transformed into IPA standard); these IPA-texts are imported into FLEx for morphological annotation. Metadata is entered by the participating researchers via the backend of the web-interface and stored in the relational database.

As mentioned above, FLEx features a build-in morphological parser, which assists the annotator with segmentation-suggestions based on prior input as well as suggestions for glossing. In OUDB, each dialect has its

own FLEx-database-file/corpus-collection; this way, we achieve a careful description of each dialect including variations (for variational linguistic purposes as well as documentation in general).

The morphological annotation layers include segmentation (see Figure 3, p. 391, layer 2), lemmas (stems and affixes; layer 3), information regarding type of variation (dialectal etc.; included in layer 3, if given), glosses (layer 4) and part-of-speech categories of stems and affixes (layer 5).

The FLEx annotated data (morphological annotated corpus data and the established stem, affix and idiom lexicon) is imported via a PHP-based data conversion and import script, which was developed in the first project phase (OUL) and has been adapted to the new requirements of the current project (OUDB), especially to the characteristics of the latest FLEx release (8.2.4). In this process, the XML-encoded FLEx export file is parsed and the retrieved lexical or textual information is imported according to the established database scheme,[5] using the unique FLEx-generated IDs as primary and foreign keys (e.g., for the connection of lexicon entries with their corresponding gloss entries, see Figure 2, overleaf).

A similar import script was developed for importing the ELAN-timecode data. Audio files are uploaded to the database together with textual metadata and an IPA transcription via the internal section.

2.2. Data Model

As said adove, the database structure for storing the FLEx corpus and lexicon data (as well as the corresponding import scripts) were already set up in the preceding project OUL and were re-used in OUDB. Figure 2 shows the representation of the data in the relational database: there is one table containing the textual metadata, one containing the IPA transcription data and two tables containing the audio data resp. the ELAN annotated audio metadata. This `elan_data` table primarily contains timecodes of the sentence boundaries as a FLEx independent representation of the audio data, allowing a sentence-by-sentence triggering of the audio recordings (see below; additional annotation is not added in ELAN.)

The FLEx annotated corpus and lexicon data are stored in several tables: an annotated token list (a segmentation of each token as well as lemma, part-of-speech tag, morpheme type and gloss of each segment) and a list of sentence translations containing the corpus data as well as

[5] In OUDB, we use a fixed set of translation languages, but the data model and the API are easily extendable to be language-independent.

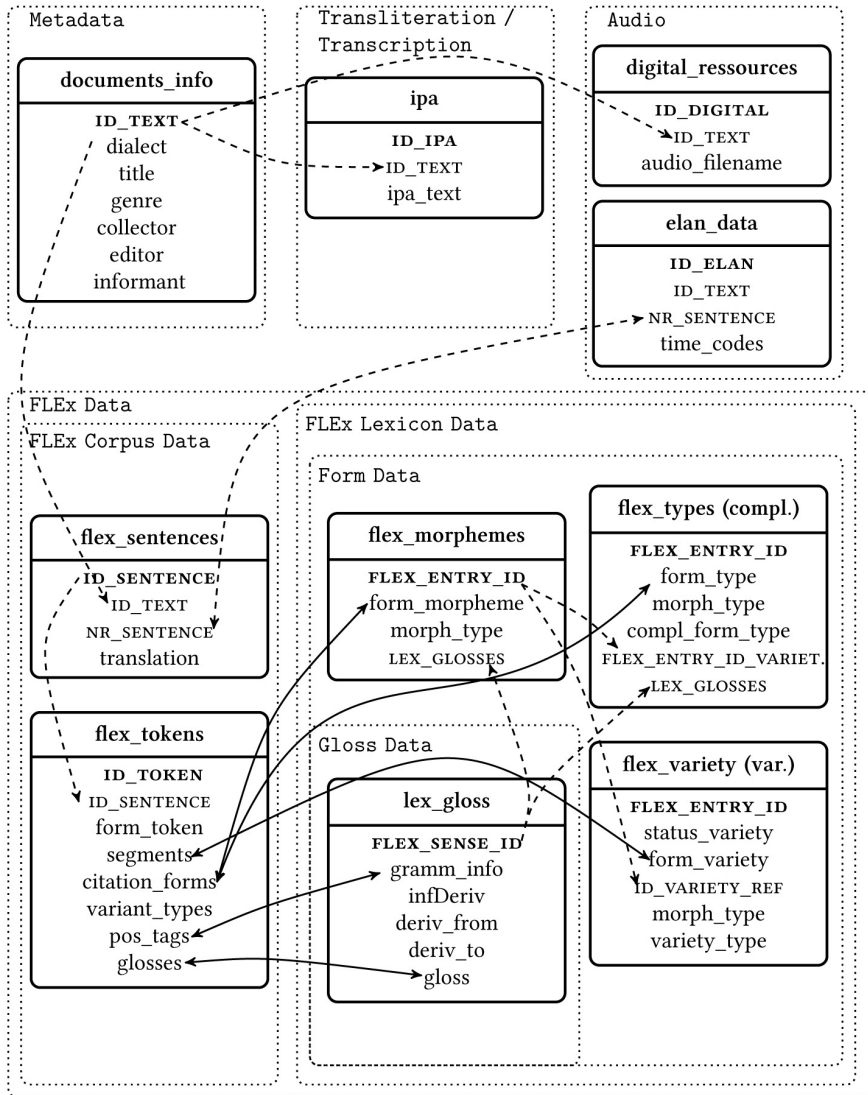

Figure 2: Relational data model for corpus, lexicon and audio data

several tables for the lexical data including stems and affixes, complex forms, variations of these primary lexicon entries and their semantic values. The aforementioned glosses are either meta-language equivalents for word stems or grammatical category labels for affixes. The foreign key

relationships between the data stored in the corresponding tables are indicated by dashed arrows in Figure 2. For instance, the corpus metadata is connected with the primary corpus data via the `flex_sentences` table based on the unique text and sentence IDs. Equally, the ELAN annotated audio data (as well as the data of the forthcoming syntactic annotation) can be connected with the FLEx data using sentence numbers, which allows a sentence-by-sentence triggering of the audio recordings from within the presentation of the glossed data.

Retrieving corresponding corpus and dictionary entries (e.g., for a concordance result of a dictionary entry) is accomplished by building ad hoc junction tables of the indexed lexicon and corpus data. The relevant columns are indexed using B-trees (Ottmann & Widmayer 1996, 317–327), allowing fast and scalable searches (McEnery & Hardie 2011, 46). In this way, the database can grow without need to change the routines and queries and the architecture of the relational database corpus arising. In general, the lexicon framework is transferable; storing the data in accordance with the relational database model keeps the data usable for later data-mining (Stonebraker & Hellerstein 2005). The multiple advantages of using relational database storage and querying for large corpora in particular are shown e.g., by Davies (2005) (cf. Gries 2009, 1237 and Gries & Berez 2017, 391); the two main advantages for OUDB are data consistency and integrity through determining constraints and scalability via relational indexing.

2.3. Annotation of information structure

Another objective of the OUDB project is to enrich the corpus data with an additional syntactic and pragmatic annotation layer, based on the principles and categories which were developed in the preceding project (OUL). For this purpose, an input form for syntactic parsing and tagging of the detected units was implemented in the existing web interface (in the interlinearized corpus view, see Figure 3, p. 391). In the glossed corpus, phrasal units are determined using a parsing algorithm implemented in the PHP-framework, based on a set of non-recursive phrase structure rules with part-of-speech-tags as terminals; some grammatical features are considered as well. Combined with an elementary clause heuristic, the identified clause and phrase chunks are subsequently checked by an annotator. The identified units as well as added analysis units (zero morphemes and possessive-suffixes for referential tagging) are in turn tagged for their functional, semantic and pragmatic role value according to the established

tagging schema using heuristics derived from linguistic regularities which are described in Janda et al. (to appear). The referential values – basis for the investigation of information structure – are to be tagged manually, but aided using a selection list of already identified referents.

This approach – utilizing an annotation form built into the existing web-based framework instead of using existing tools – was chosen, because, this way, it was possible to combine a shallow parsing, identifying the basic syntactic units,[6] with manual correction and immediate tagging of the parsed units and to write the gained data directly in the database, connected with the initial flex data, therefore avoiding repeated importing and exporting operations and the development of the necessary conversion tools. Also, this way, using the well-known display of the data in the research environment for further annotation, it is not necessary for the participating researchers to become familiar with new interfaces (cf. Black & Simons 2006, 39ff); moreover, the development of the phrase structure grammar as well as the set of tagging rules could be performed gradually, expanding and adapting the rules without the need to constantly export the still growing corpus for parsing.

3. Results

3.1. Data output

In OUDB, we re-used and expanded the already existing web interface (developed in the preceding project OUL, including menus in English, Russian and German) for online access to the database via corpus and lexicon modules. There are two ways to access the corpus data via the OUDB website: the "Text Corpus" section (where the texts are available according to their metadata) and the "Concordance" section. In using the concordance module to generate a lexicon entry-specific concordance, the corpus-based dictionary provides alternative access to the corpus data in addition to the concordance interface. As the concordance and the lexicon also allow input in Cyrillic, and as the lexicon features glosses in Russian, the lexicon and corpus becomes accessible to members of the language community as well.

[6] The only existing parsers for Khanty and Mansi in the *giellatekno*-framework being only in a rudimentary state, featuring only some test data, thus were not sufficient for the task.

3.2. Corpus module

The glossed corpus data is compiled and displayed on the website sentence-by-sentence in an interlinearized display style following the Leipzig Glossing Rules[7] including layers giving the lemmatization (layer 3) and part-of-speech data (layer 5) next to the glossing-layer (4). There are English, German, Russian and Hungarian translations, if available. Each token and sentence is accessible by its ID, which is used to connect a concordance search result with the glossed text and to highlight the relevant token(s) (see Figure 3).

jɐ:	#,#	tʲu:	i:kinə	ɬi:totɐt-qu:ɬɐt	ɬi:pti	#.#
jɐ:		tʲu:	i:ki-nə	ɬi:tot-ɐt qu:ɬ-ɐt	ɬi:pt-i	
jɐ:		tʲu:	i:ki-nə	ɬi:tot-ɐt qu:ɬ-ɐt	ɬɐ:pət+[PST]-i	
well		that	old_man-LOC	food-INSC fish-INSC	feed+[PST]-PASS.3SG	
ptcl		dem.dist	subs-infl:n	subs-infl:n subs-infl:n	v-infl:v	

So, the old man gave him some food and fish to eat.
Ну, старик угостил молодого человека едой-рыбой.

Figure 3: Passive construction with locative coded agentive-like argument

3.3. Concordance module

The two main control elements of the concordance web interface (see Figure 4, overleaf) are the *search bar* with an input field and drop-down menus, which allows the user to filter and sort the search results, and the *IPA input toolbar*. This virtual keyboard allows users to enter IPA characters (client side processed via javascript), and also serves as a matching chart for a fuzzy search within the corpus, using ASCII characters as cover symbols for defined IPA character classes (see Figure 4). For matching classes of Ob-Ugric IPA characters with ASCII characters, we use an associative array as data structure with the ASCII cover symbols as keys for and arrays of the matching IPA symbols as related values, which are subsequently used in regular expressions within the SQL queries. The results can be

[7] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

sorted in numerous ways, including a reverse alphabetical ordering of the left context (right-to-left).

Newly, the input of Cyrillic letters in the fuzzy search mode is possible as well, translating these in corresponding IPA-letters.

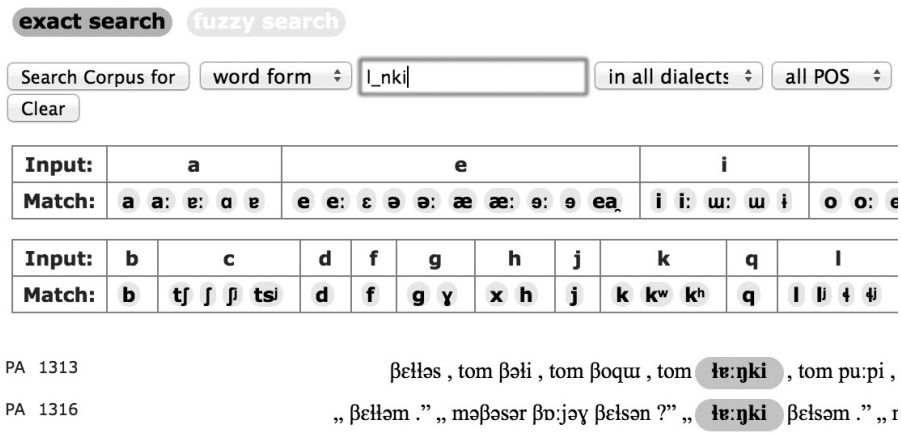

Figure 4: Details of a concordance search

The corpus is searchable for lemmas (stems and affixes), word forms and glosses. It is possible to specify the part-of-speech category of the token in search; wildcards (∗ or % for an unspecified number of characters, _ for exactly one single character) can be used as well. Regular expressions in queries can be used to search for word forms and lemmas.

Each result of a concordance query is linked with the corresponding location in the corpus, where the relevant token is highlighted (see Figure 3).

The *multiple glosses* search option expands the search from one token form or its gloss (or the glosses of its individual morphemes) to more detailed searches for multiple values in one token or values in different tokens and thus offers the possibility of searching syntactic patterns. The user enters a string with two arguments (the two search terms), whereas the optional third argument specifies the window size; without specification, the standard search radius is sentence-wide. There is an “exact” option, which restricts the search to the given distance of the two tokens instead of a search window of the given size. There is also a “left/right” option, which takes the order of elements into consideration. Combined with the wildcard % and the part-of-speech restriction for the base token (first argument), advanced and versatile queries are possible, e.g., a search

for morphosyntactic patterns such as specific preverbal or postpositional constructions (cf. Bradley 2015; Schön 2015):

1. `% PTCP.PRS 1`, pos=`preverb` + `right` → *preverbal present participle construction*[8]
2. `% PTCP% 1`, pos=`pstp` + `left` → *postpositional participle construction*
3. `%DAT% PASS%`, pos=`ppron` → *passive construction featuring a pronominal indirect object* (window-size=`sentence`)
4. `LOC PASS% 2`, pos=`subs` + `right` → *passive construction with locative coded agentive-like argument following immediately or with distance ≤ 2 from the verb*, cf. Filtchenko (2006), see Figure 3.

A search for the occurrence of two different glosses in the same token is possible as well, namely by defining a window size of `0`. This way, in combination with a wildcard, the concordance cannot only be used to search for a specific form or gloss (or a combination of these), but for all occurrences of a part-of-speech category:

1. `%SG% LOC 0` → *morpheme chain with any singular possessive suffix and a locative case suffix*
2. `% % 0`, pos=`prvb` → *complete concordance of the preverbs in the corpus.*

As shown, the indexed, semi-automatically annotated (and thus very accurate) corpus data can be used, e.g., to perform complex constructional pattern queries, which are suitable for tackling advanced morphosyntactic questions.

3.4. Dictionary module

Just as the concordance module, the dictionary section uses the IPA input toolbar and features a fuzzy search with ASCII cover symbols within the lexical data (stems, affixes and their english or russian glosses). As already mentioned, the lexical data is linked with the corpus data (which it is

[8] For the given corpus, the queries show a good performance. For instance it takes 75 ms runtime for the query for preverbal present participle constructions (see above; corpus size at the time of measurement: 30,000 tokens).

based on) by the concordance module, which is incorporated in the view of a lexicon entry.

Vice versa, the glosses in the interlinear corpus view are linked to a search query for the gloss in the dictionary.

3.5. Audio section

As mentioned above, the ELAN annotated audio timecode data is used for a sentence-by-sentence triggering of the audio recordings via javascript, thus interactively connecting transcription and audio data. The planned integration of this functionality in the display of the glossed corpus data, which is possible by exploiting the unified data model (see above), will lead to a growing connection of the audio and transcription data with the multiple layers of annotational data.

3.6. Exporting and archiving

One of the main objectives of OUDB is to provide not only a comprehensive access to the data via the connected web modules, but also to provide the data itself in various formats for multiple purposes. Next to an export function of sentences as language examples (in HTML and TeX), exports of concordance query results in relational format and the possibility to bookmark the result of a concordance query, the complete data will be made available for download at the end of the project, along with the developed data processing scripts (via GitHub, s. below).

This will include a download of the complete SQL database as well as files in ELAN-XML together with the corresponding audio recordings (if present) and metadata in CMDI-XML-format, both of which are used for archiving the corpus data in *The Language Archive*[9] at the Max Planck Institute for Psychlinguistics; the creation of the OUDB archive[10] is carried out in collaboration with the *Langdoc Group*[11] (University of Freiburg) and the OUDB-corpus will be part of the *Permic-Varieties* corpus in TLA. The necessary data conversion tools are currently being developed and are documented and retrievable via GitHub.[12]

[9] https://tla.mpi.nl

[10] https://hdl.handle.net/1839/88A38A1F-5367-4415-B5F9-E86E231AFE66@view

[11] http://langdoc.github.io

[12] https://github.com/langdoc/OUDB

4. Conclusion

As outlined in this paper, *Ob-Ugric database* (OUDB) is a web-based framework for the storage and advanced retrieval of annotated corpora and corpus-based lexical databases of Khanty and Mansi dialects, that makes diversely annotated text and audio corpora of these less described Ob-Ugric dialects available and accessible in a unified database. The provided, richly annotated language resources not only serves as a usage-based description for language documentation, but also as research material for typologists and variational or cognitive linguists as well as cultural anthropologists. Furthermore, as discussed in section 1, footnote 2, such small core corpora of not yet or less described languages may as well function as training data for probabilistic tagging or parsing and thus, in helping building language technology resources, may play a vital part in language revitalization.

The corpus building has been carried out utilizing established documentation and annotation tools; due to the small corpus size and the heterogeneity of the data, a semi-automatic approach for the morphological annotation was chosen in the preceding project OUL and has been retained for OUDB, using the annotation tool FLEx, which automatically generates a lexicon based on the annotated corpus data. Imported in a relational database, these corpora and corpus-based lexicon data are then used for output and further processing via web interface as well as for providing the raw data for archiving matters as well as corpus linguistic research.

The developed online tools give straightforward access to the data; they include intertwined concordance and dictionary modules, which are designed to be used not only by researchers but by the members of the language community as well, featuring a fuzzy search and Cyrillic input possibility.

At the end of the project, the complete database will be made available for download in relational format via the OUDB-website; the data will also be archived in an ELAN-XML-format at *The Language Archive* as part of the *Permic-Varieties* corpus (see above), allowing the use of the corpus data in future research.

OUDB can be considered as part of a greater research program which aims to provide and share corpus data in a standardized way and builds on extensive annotation as a way of enriching the primary speech data, thus allowing sophisticated investigation of patterns of language use on different levels of linguistic description.

References

Black, H. Andrew and Gary F. Simons. 2006. The SIL FieldWorks Language Explorer approach to morphological parsing. In Computational Linguistics for Less-Studied Languages: Proceedings of Texas Linguistics Society 10. Austin, TX: CSLI Publications. 37–55.

Bradley, Jeremy. 2015. Corpus.mari-language.com: A rudimentary corpus searchable by syntactic and morphological patterns. In First International Workshop on Computational Linguistics for Uralic Languages. 57–68.

Davies, Mark. 2005. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. International Journal of Corpus Linguistics 10. 307–334.

Filtchenko, Andrey. 2006. The Eastern Khanty locative-agent constructions. In B. Lyngfelt and T. Solstadt (eds.) Demoting the Agent: Passive, Middle and Other Voice Phenomena. Amsterdam & Philadelphia: John Benjamins. 47–82.

Gries, Stefan Th. 2009. What is corpus linguistics? Language and Linguistics Compass 3. 1225–1241.

Gries, Stefan Th. and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In N. Ide and J. Pustejovsky (eds.) Handbook of Linguistic Annotation. Berlin & New York: Springer. 379–409.

Hardie, Andrew. 2012. CQPweb – Combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics 17. 380–409.

Janda, Gwen Eva, Axel Wisiorek and Stefanie Eckmann. to appear. Reference tracking mechanisms and automatic annotation based on Ob-Ugric information structure. Journal de la Société Finno-Ougrienne 96.

McEnery, Tony and Andrew Hardie. 2011. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

Ottmann, Thomas and Peter Widmayer. 1996. Algorithmen und Datenstrukturen. Heidelberg: Spektrum.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing. Manchester. 44–49.

Schön, Zsófia. 2015. On the road to a dialect dictionary of Khanty postpositions. In First International Workshop on Computational Linguistics for Uralic Languages. 99–107.

Stonebraker, Michael and Joey Hellerstein. 2005. What goes around comes around. In J. Hellerstein and M. Stonebraker (eds.) Readings in Database Systems. Cambridge, MA: MIT Press. 2–41.