Obugric Database: 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

December 31, 2015

Abstract

This paper aims to present a comprehensive web-based framework for the storage and advanced retrieval of annotated corpora and corpus-based lexical databases of Khanty and Mansi dialects within the framework of the project *Ob-Ugric database: analysed text corpora and dictionaries for less described Ob-Ugric dialects* (OUDB). The strength of this approach lies in combining semi-automatic annotation using established documentation and analysis tools with modern web technologies and relational databases.

Key aspects are: Extensive annotation, which covers different levels of linguistic description as well as language internal variation; performing intricate concordance searches based on the annotational linguistic metadata, using a well-adapted relational database scheme that allows complex but nonetheless fast and scalable queries over indexed data; making it possible to identify not only single token forms but *constructional patterns* on various linguistic levels, allowing cutting-edge usage-based research including new corpus evaluation methods

such as 'collostructional' analysis; offering a web interface which provides comprehensive access to the corpus and lexicon data from any up-to-date browser, the client-server framework guaranteeing platform independency; establishing a collaborative research platform with a differentiated user management system which enables contributing researchers to upload their material to the database; providing output that conforms to linguistic standards that is simultaneously suitable as an export format for sharing and archiving data.

As OUDB is work in progress, not all of these features have been fully implemented yet, but the main functionality of the projected framework is existent and operational.

1 Introduction

The project *Ob-Ugric database: analysed text corpora and dictionaries for less described Ob-Ugric dialects* (OUDB, since July 2014)[1] and its framework presented in this paper focus, among other things, on developing semi-automatically tagged corpora and lexical databases for dialects of the Khanty and Mansi languages, belonging to the Ob-Ugric branch of the Finno-Ugric language family. Currently, the size of the glossed corpus is about 30,000 tokens, with the total corpus having over 200,000 tokens in approximately 400 texts in IPA transliteration/transcription.

The corpora and databases were initially set up in the course of the project *Ob-Ugric languages: conceptual structures, lexicon, constructions, categories* (BABEL, August 2009–July 2012), which contained two Khanty (Kazym and Surgut) and two Mansi (Northern and Southern) dialects. As this initial project of the universities of Munich, Vienna, Szeged and Helsinki primarily dealt with already published written material, the documentation and analysis software FieldWorks Language Explorer (FLEx)[2] was chosen for the data analysis, which proposes annotations based on the prior input.

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 by only text editions from the end of the 19th century, the Yugan dialect of Khanty mostly relies on transcribed sound recordings from fieldwork in the 21st century. To accomodate this circumstance, the annotation tool ELAN was added to our tool set for data handling.

[1] http://www.oudb.gwi.uni-muenchen.de/
[2] http://fieldworks.sil.org/flex

2 Technological Framework

OUDB is hosted and maintained by the IT Group for the Humanities of the LMU Munich (ITG), which offers an Apache web server as well as a MySQL server, thus providing a perfect environment for establishing a web-based research platform such as OUDB. Main advantages of this client-server model are platform independency, long-term availability and easy international collaboration [1, 2, p. 45 ff.]. The fundamental database structure and the PHP-based website (including a backend for cooperating researchers) were established in the first phase of the project (BABEL). On this basis, OUDB continues to develop advanced corpus and lexicon tools[3], with expanded filter possibilities, a new interface, faster and more complex queries, 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 (phonetic, morphologic, syntactic, semantic, pragmatic), as well as a corpus-based electronic dictionary connected with the concordance module. The following presentation will mainly focus on the database representation of the annotated corpus data and the characteristics of the concordance search.

3 Structure of the Database

3.1 Importing the Data

Audio files are uploaded to the database, together with textual metadata and an IPA transliteration/transcription via the internal section. Each database entry is indexed in the process. The FLEx annotated data is imported via a stand-alone PHP script originally written by Susanne Grandmontagne (ITG) in the first project phase (BABEL) and adapted to the new requirements of the current project, especially to the characteristics of the latest FLEx release (8.2.4.). The XML-encoded FLEx export file is parsed and the lexical or textual information retrieved in the process is imported according to the established database scheme, using the unique flex-generated IDs as primary and foreign keys.

3.2 Data Scheme

Figure 1 shows the representation of the data in the relational database: there is one table containing the textual metadata, one containing the IPA transliteration/transcription

[3]Tools will be provided by the authors on request and are envisaged to be published at completion of the project under a Creative Commons Licence.

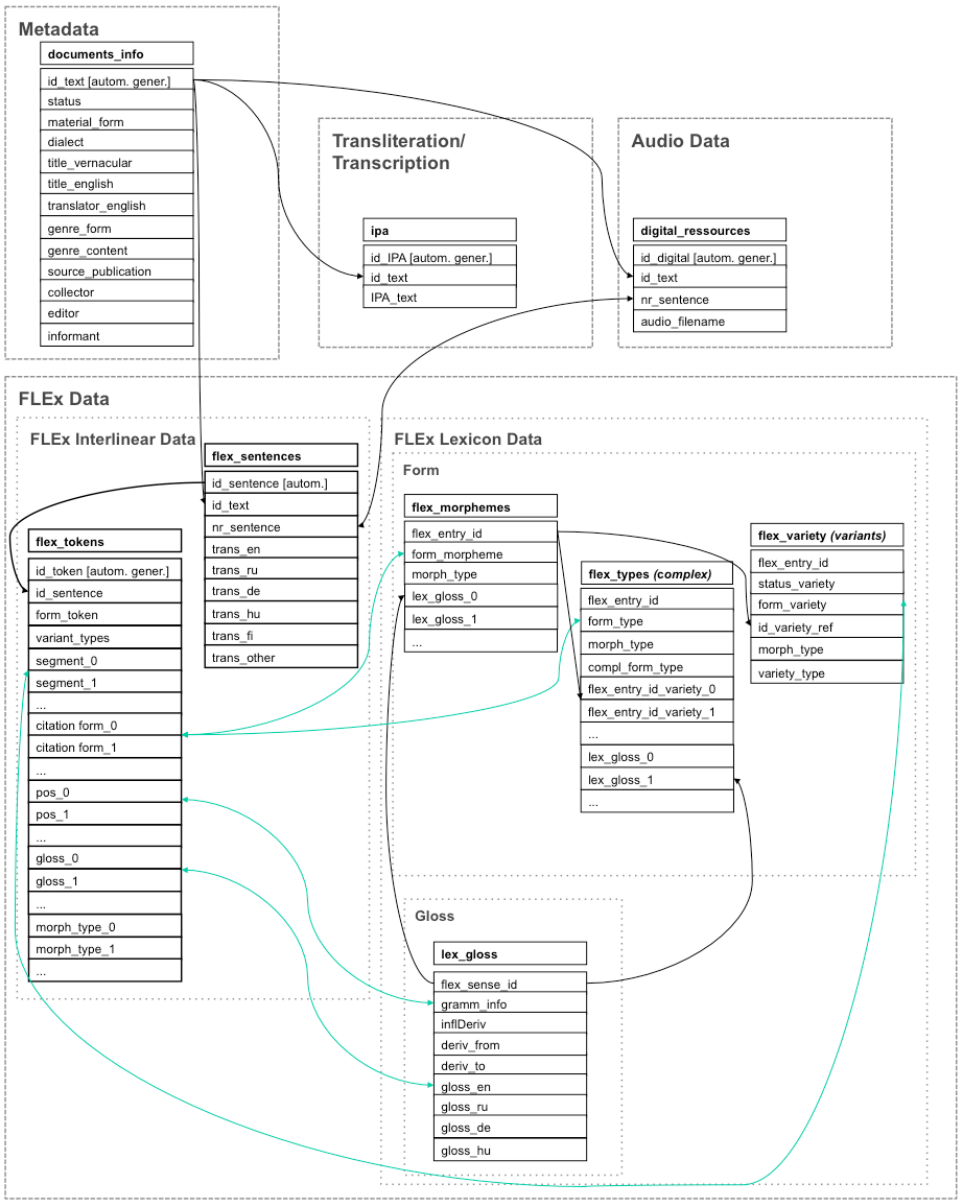

Figure 1: Representation of corpus and lexicon data in the database

data and one containing the audio data. The FLEx annotated corpus and lexicon data are stored in several tables: an annotated token list (a segmentation of each token as well as the citation form, part of speech tag, morpheme type and gloss of each segment) and a list of sentence translations containing the corpus data as well as several tables for the lexical data including morphemes, 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 black arrows in Figure 1. 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. In a further step the ELAN annotated audio data will be connected with the FLEx data using sentence numbers, which will allow a sentence by sentence triggering of the audio recordings via javascript[4].

As FLEx does not offer the possibility to export text and lexicon data in combination, the information on the relationships between the corpus and the corpus-based lexicon data (indicated by blue arrows in Figure 1) is not part of the imported data. Retrieving corresponding corpus and dictionary entries (e.g. for a concordance result of a dictionary entry) is therefore accomplished by building ad hoc junction tables of the indexed lexicon and corpus data. The relevant columns are indexed using B-trees [3, p. 317–327], allowing fast and scalable searches [1, p. 46]. Like this, the database can grow without the need to change the routines and queries and the architecture of the relational database corpus arising. The lexicon framework is transferable in principle; storing the data in accordance with the relational database model keeps the data usable for later data-mining [4]. The multiple advantages of using relational database storage and querying for large corpora in particular are shown e.g. by Davies [5] (cf. [6, p. 13] and [7, p. 13]); the two main advantages for OUDB are data consistency/integrity through determining constraints and scalability through relational indexing.

4 Analyzing the Data

4.1 User Interface

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 (which the following description will be about).

[4]View Text 'piːtʲəŋkəliɣən-oːpisɐɣən A' (ID 732, Surgut Khanty), „Audio + Metadata"; this tool can be adapted for video files as well since it uses HTML 5 standard elements.

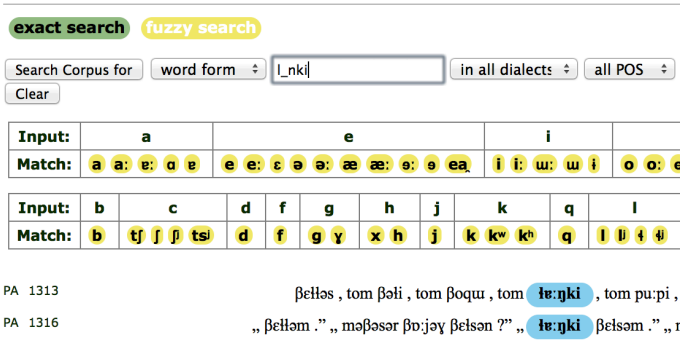

Figure 2: Details of a concordance search

The two main control elements of the concordance web interface (see Figure 2) 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 2). For matching classes of Ob-Ugric IPA characters with ASCII characters, we use an associative array as a data structure with the ASCII cover symbols as keys and arrays of the matching IPA symbols as related values. We make heavy use of regular expressions within the SQL queries. The results can be sorted in numerous ways, including a reverse alphabetical ordering of the left context (right-to-left). The principal sorting order for the IPA transliteration/transcription of Ob-Ugric languages is implemented in the SQL queries using a sorting array[5].

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.

4.2 Querying the Corpus Data

Corpus searches rely on SQL as query language. Our framework makes use of the data relations applied in the database scheme to retrieve the relevant tokens. The context of a token is retrieved by multiple self-joins of the token table using the token IDs. 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 corpus is

[5]E.g. reverse sorting of the left context in a KWIC: `FIELD(left(reverse(lk),1),$alph)`.

searchable for word forms, morphemes (stems and affixes) 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 lemmata.

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[6]. 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 morphosyntactical patterns such as specific preverbal or postpositional constructions, cf. [8, 9]:

1. `% PTCP.PRS 1`, pos=`preverb` + `right` ⟶ *preverbal present participle construction*
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. [10], see Figure 3.

A search for the occurrence of two different glosses in the same token is possible as well, by defining a window size of `0`. This way, in combination with a wildcard, the concordance can not 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 corpus.*

[6] This search type uses self-joins of the token list according to established criteria, e.g. `join on (t1.id_token = t2.id_token-1 OR t2.id_token = t1.id_token-1)` for a search window size of 1 or `join on t1.id_sentence = t2.id_sentence` for a sentence wide search, and complex `where`-restrictions using joins on the metadata, e.g. `where (t1.gls_0 LIKE 'squirrel' OR t1.gls_1 LIKE 'squirrel' ...) AND (t2.gls_0 LIKE 'LOC' OR t2.gls_1 LIKE 'LOC' ...)`.

jɐ:	#,#	ʧu:	i:kinə	ɬi:totɐt-qu:ɬɐt	ɬi:pti	#.#
jɐ:		ʧu:	i:ki-nə	ɬi:tot-ɐt qu:ɬ-ɐt	ɬi:pt-i	
jɐ:		ʧu:	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.
Hát, az öreg étellel-hallal megette.

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

This presented search syntax is only a sketch of what will follow. It will be expanded and universalized, allowing the definition of window size and part-of-speech categories directly in the input and keeping existing query syntaxes like BNC or CQP in mind (cf. [5] and [2]). Our main goal will be the expansion of this multiple gloss search framework to a generalized construction search framework in which each base token of a construction represents this construction (as its head) and can be recursively be part of a bigger construction, establishing a free morphosyntactic constructional search syntax that will be much more adaptable than a linear selection of categories e.g. via selection menus. This expanded search functionality will feature nested queries, each subquery embodied by bracketing and corresponding in principle one binary *multiple glosses* SQL query as shown above, where each base token will function as an identifier for each sub-construction in the complex construction query. Here are two examples for possible nested construction queries:

1. `((%=v %=prvb)1-left %LOC%=ppron)clause` ⟶ *a clause represented by a verb phrase featuring a preverb and a locative coded pronoun*

2. `(PST=v (%=pstp PTCP.PRS=v)clause)sentence` ⟶ *a complex monofinite sentence construction featuring an anteriority postpositional participle construction.*

Exploiting the multilayered, structured representation of the linear speech data in the relational database (e.g. clause/sentence IDs in combination with token IDs), it becomes possible to express a combination of morphologic, syntactic as well as pragmatic or semantic features in one query, forming a complex linguistic pattern and displaying this construction in context. For the given corpus of about 30,000 tokens, the queries show a good performance[7].

[7]For instance it takes 75 ms runtime for the query for preverbal present participle constructions (see above). As the OUDB framework is developed primarily as an integrated research environment connecting

5 Output of Data

The glossed corpus data is compiled and displayed on the website sentence-by-sentence in an interlinearized display style following the Leipzig Glossing Rules [11], with additional lemmatization and part-of-speech data, including English, German, Russian and Hungarian translations. Each token and sentence is accessible by its ID, which is used to connect a KWIC result with the glossed text and to highlight the relevant token(s) (see Figure 3).

6 Future Goals

As OUDB is work in progress, there will be a constant expansion of the range of functionalities offered by our frameworks. As regards the corpus, there will be two main updates. Firstly, an export tool for the preparation of structured data for client-based evaluation as well as for possible archiving, and the accompanying XML output implementation, will be realized. Secondly, we will develop a syntactic and pragmatic annotation system compatible with our existing database scheme. This forthcoming semi-automatic annotation tool, which is already rudimentary implemented, will use the existing FLEx data (esp. part of speech data) for providing a parenthesized annotation line of each sentence using constituent analysis rules. This annotation line can be manually checked and complemented by the annotator with additional syntactic and pragmatic tags as well as additional levels of syntactic analysis (clause). The parenthesized annotation data is then saved in an extra table in the database and simultaneously parsed in a multidimensional array[8], which is used to update the entries in the `flex_tokens` table with their corresponding syntactic and pragmatic annotations. These additional layers of annotation (which will be included in the interlinearized presentation of the corpus)[9] expand the search functionality for constructions even further. Through providing a clause-specific search window, a much more precise identification of syntactical patterns will be possible.

Regarding the concordancer, we are planning an extension which will enable ad-

corpus, lexicon and audio data of the small heterogeneous corpora of the Ob-Ugric languages (e.g. including language specific IPA-ASCII-translation rules in the corpus and lexicon search tools), the application for bigger corpora is not main objective, but we are generally working on improving the performance through extended indexing and enhanced queries on the basis of which the applicability for larger corpora will be evaluated.

[8] This php parsing module will equally be used in the intended construction query system.

[9] In this context, a script for the online visualization of syntactic trees developed at the ITG (LMU Munich) for the *Biblia Hebraica transcripta* (Richter, Eckardt, Specht, Argenton, Zirkel, Riepl, Teuber) will be adapted.

vanced statistical testing. As the basic implementation of a collocational analysis is already implemented in the concordancer with the *multiple glosses* option (see above), the frequency data of these query results can easily be obtained and processed with statistical algorithms, also incorporating measurements of effect size. Thanks to the (already implemented, and in the future expanded) construction search functionality this framework is especially suitable for new construction-based corpus analysis methods such as the 'collostructional' analysis, a constructional grammar-based extension of collocational analysis proposed by Stefanowitsch and Gries [12] where the p-values of a Fisher's exact test resp. the odds ratio are used as a measure of the association strength of a lexeme in a construction. We will be looking into the possibility of using n-gram frequency tables (resp. views) as proposed by Davies [5] for faster collocational analysis, as well as possible construction tables, building a kind of 'constructicon' [13], e.g. containing frequency information of lexical units concerning a certain slot of a construction.

7 Conclusion

As outlined in this paper, OUDB aims to give researchers around the world a server-based – thus client-independent – corpus and lexicon tool that will make corpora of the less described Ob-Ugric dialects available and accessible in connection with lexical and audio data. Thus, this multipurpose corpus data will serve not only language documentation [6, p. 13 f.], but can also serve as research material for typologists and variational or cognitive linguists. In using free Software such as MySQL and PHP, the framework we developed imposes no restrictions on providing and sharing modules.

Using the indexed, semi-automatically annotated (and thus very accurate) corpus data, complex constructional pattern queries are possible, allowing users to tackle advanced morphosyntactic questions. Through the planned standard format export function, researchers will be able to retrieve data for their own evaluation (using R, Perl etc.). OUDB can be considered 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 linguistic investigation of complex patterns of language use.

References

[1] Tony McEnery and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice.* Cambridge Textbooks in Linguistics. Cambridge University Press, 2011.

[2] Andrew Hardie. CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3):380–409, 2012.

[3] Thomas Ottmann and Peter Widmayer. *Algorithmen und Datenstrukturen*. Spektrum, Heidelberg, 1996.

[4] Michael Stonebraker and Joey Hellerstein. What goes around comes around. *Readings in Database Systems*, 4, 2005.

[5] Mark Davies. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, 10(3):307–334, 2005.

[6] Stefan Th. Gries. What is corpus linguistics? *Language and linguistics compass*, 3(5):1–17, 2009.

[7] Stefan Th. Gries and Andrea L. Berez. Linguistic annotation in/for corpus linguistics. http://www.linguistics.ucsb.edu/faculty/stgries/research/InProgr_STG_alb_lingannotcorpling_hboflingannot.pdf, September 2015.

[8] Zsófia Schön. On the Road to a Dialect Dictionary of Khanty Postpositions. In *Septentrio Conference Series*, pages 99–107, 2015.

[9] Jeremy Bradley. Corpus. mari-language. com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns. In *Septentrio Conference Series*, pages 57–68, 2015.

[10] Andrey Filtchenko. The Eastern Khanty locative-agent constructions. In *Demoting the Agent: Passive, Middle and Other Voice Phenomena*, pages 47–82. 2006.

[11] Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. The Leipzig Glossing Rules. Conventions for interlinear morpheme by morpheme glosses. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf, November 2015.

[12] Anatol Stefanowitsch and Stefan Th. Gries. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243, 2003.

[13] Charles J. Fillmore. Border conflicts: FrameNet meets construction grammar. In *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*, pages 49–68, 2008.