Languages under the influence: Building a database of Uralic languages

Eszter Simon, Nikolett Mus
Research Institute for Linguistics
Hungarian Academy of Sciences
{simon.eszter, mus.nikolett}@nytud.mta.hu

Abstract

For most of the Uralic languages, there is a lack of systematically collected, consequently transcribed and morphologically annotated text corpora. This paper sums up the steps, the preliminary results and the future directions of building a linguistic corpus of some Uralic languages, namely Tundra Nenets, Udmurt, Synya Khanty, and Surgut Khanty. The experiences of building a corpus containing both old and modern, and written and oral data samples are discussed. Principles concerning data collection strategies of languages with different level of vitality and endangerment are discussed. Methodologies and challenges of data processing, and the levels of linguistic annotation are also described in detail.

1 Introduction

This paper sums up the steps, the preliminary results and the future directions of building a linguistic corpus of some Uralic languages within the research project called *Languages under the Influence*. The project started in February 2016 and lasts until July 2017 and is funded by the Hungarian National Research, Development and Innovation Office (grant ID: ERC_HU_15 118079).

It is a pre-ERC project getting national support to enter the European Research Council (ERC) programme[1], thus it is a pilot project. Its aim is to create the theoretical

[1] https://erc.europa.eu/

The 3rd International Workshop for Computational Linguistics of Uralic Languages, pages 10–24,
St. Petersburg, Russia, 23–24 January 2017. ©2017 Association for Computational Linguistics

and methodological basis of an ERC project proposal, which is called as 'the main project' in the article.

The main project has a twofold objective. As for the theoretical investigations, it focuses on potential syntactic changes affected by the heavy influence of the Russian language in several Uralic languages spoken in the Russian Federation (mainly in Siberia). Our research covers, among others, the change of basic word order (i.e. the SOV-to-SVO order), the spreading of finite subordination and clause-initial complementizers, and the diversification of indefinite pronouns. The languages involved are *Tundra Nenets*, *Udmurt*, *Synya Khanty*, and *Surgut Khanty*.

On the other hand, the main project's computational objective is to build a linguistically annotated database of written and spoken sources in the aforementioned languages, which makes it possible to research on Uralic–Russian language contacts. In order to observe syntactic changes of minority languages under the influence, we aim at processing texts collected from different times. It is known, however, that the literacy of Uralic languages do not have a long history and tradition. The oldest sources included in our project originate from the beginning of the 20th century (when organized expeditions were undertaken in order to document and describe Uralic languages). In addition, we gather data from published and/or electronically accessible sources, which are presented by fieldworks undertaken recently. We focus on selecting texts provided by as many authors as possible from different social classes, age, sex, dialects and genres. Furthermore, our activities will include fieldwork, during which we will collect contemporary spoken language material, thus the database will represent the written and the spoken versions of the languages as well.

Within the pilot project, text samples from the two time periods for all the aforementioned languages have been collected and partly processed and annotated. In this paper, we discuss considerations taken into account and methods used in the pilot corpus building process, which indeed will be applied in the main project as well. The structure of the remaining part of the article is as follows: In Section 2, we discuss the main principal criteria behind corpus building. Section 3 describes the collected text material and the problems we face with when aiming at creating such a corpus. The text processing steps are detailed in Section 4, while Section 5 presents the structure of the corpus. We conclude in Section 6, which contains the future directions as well.

2 Theoretical considerations

We follow several principal criteria within the pilot project which form the basis even of the whole corpus building process of the main project. Since the Uralic languages dealt with are endangered and/or poorly documented, we think that creating a database which follows the basic concepts of language documentation is of huge importance, as detailed in 2.1. Other principal criteria are: following international standards on every level of corpus building (2.2), consistency which would apply on all levels (2.3), and using and creating freely available resources (2.4).

2.1 Language documentation

One of the principal criteria of the pilot project is following some basic principles of language documentation (cf. [1, 2]). First, we focus on compiling *primary data* in Himmelmann's sense [3], i.e. we collected data types that were produced at a specific point in time and space by a specific speaker instead of collecting and using generalized secondary data, e.g. elicited data (that may be corrupted). This rule is also applied to our fieldwork(s), during which we prefer to document the languages in their natural forms, i.e. to record language variants that are not influenced by the style of genres (as in the case of folklore texts), or by some prescriptive considerations (as in the case of journalism). The application of this principle led us to the second concept, i.e. collecting and storing the *context* of the data (metadata), which generally concerns the recording time and place, the age, gender and spoken dialects of the informants. The collection of metadata is also needed for other scientific fields, e.g. for sociolinguistics, anthropology, and sociology [4]. Third, we aim at building a database whose content is *transferable* both in linguistic and in technical terms. Consequently, both the representation, i.e. the transcription, and the analysis, i.e. the morphological tags, of our data is not restricted to any theory or method. Finally, we make our data *available* for further theoretical and applied research, as well as for direct use by the relevant language communities.

However, certain rules and conditions of language documentation are necessarily contravened in our project. For instance, as languages are still most typically used in speech, collecting spoken data has primacy in language documentation. This principle, however, has its limits in the case of the historical data, such as the availability of the original sound recordings. Therefore, we decided to collect old texts that appeared in critical editions instead of inscriptions or original manuscripts. Although these editions usually contain smaller range of text genre, they still provide rich metadata and useful information about the rules followed during the transcription of spoken data. In the case of contemporary written language samples, we aim for selecting texts which are provided by as many authors as possible from different social classes, age, gender, dialects and genres. As our project has its time limit, we need to break the rule of collecting full range of textual genres and registers and primarily focus on those ones which may typically show the results of a potential language contact. These written genres are usually the types which are closer to the spoken language, such as blogs and tweets, see more details in Section 3. Finally, the written language varieties of the languages concerned are typically produced by only a few writers, therefore, the sociolinguistical parameters may not be balanced in our corpus.

2.2 International standards

Another criterium is following international standards on every level of the corpus building process. Therefore, we only use standard Unicode characters, we provide phonemic transcription using the letters of the International Phonetic Alphabet (IPA), we follow the Leipzig Glossing Rules (LGR)[2] and abbreviations, and we use standard file formats, as detailed below.

The Unicode Standard[3] is a multilingual coding system which provides a consistent encoding for most of the world's writing systems. Recently, it became an international standard, which supports the worldwide interchange, processing and display of written texts of diverse languages. One of the great advantages of Unicode is that it properly handles various accented and multi-accented characters, since basic characters and combining diacritical marks are represented by their own codes. Unicode also contains all of the Cyrillic characters used in the orthographies of the aforementioned languages. Moreover, the Unicode Consortium provides new supplements in each release for users to be able to handle and represent the proper characters used by minority communities. For example, the characters Ԯ (U+052E Cyrillic capital letter el with descender) and ԯ (U+052F Cyrillic small letter el with descender) were released in version 7.0 in 2014 based on the character request proposal submitted to the Unicode Technical Committee by Tapani Salminen[4]. Salminen provides evidence from recent native publications that these characters are needed, since a descender is used in common typographic practice of the Northern Khanty, Eastern Khanty, Tundra Nenets and Forest Nenets languages. Before version 7.0, replacement characters were used instead of them, mostly Ԓ (U+0512 Cyrillic capital letter el with hook) and ԓ (U+0513 Cyrillic small letter el with hook) or Ӆ (U+04C5 Cyrillic capital letter el with tail) and ӆ (U+04C6 Cyrillic small letter el with tail), but since then the proper characters be can used.

The Unicode code charts also contain all of the widely used IPA characters, thus every textual element (Uralic transcriptions, IPA transliteration, Cyrillic characters) can be stored with standard Unicode characters, which makes it possible to replace the old, makeshift font collections and to follow the international standards.

[2] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

[3] http://unicode.org/

[4] http://www.unicode.org/L2/L2012/12052-khanty-nenets.pdf

Language documenters of the languages concerned in this project used different subtypes of the traditional Finno-Ugric transcription (FUT) system (see details in Section 4.2). However, these Latin-based transcription systems are not standardized nor unified, even within one language. For this reason, it is important to publish the texts using the IPA system, as it makes the texts readable for further areas of linguistics outside Uralistics.

On the morphological annotation level, we follow the Leipzig Glossing Rules with some modifications. The tokens and the corresponding pieces of morphological information are aligned, i.e. the following annotations are added to each token of the corpus: lemma, part-of-speech (POS) tag, morphological labels of derivational and inflectional categories, and English translation of the lemma. The glosses are converted from the output of a morphological analyzer developed for the language concerned. As a consequence, if the output is not segmented on the level of morphemes, even the glosses will not contain morpheme-by-morpheme correspondence. In the case if the morphological analyzer is able to produce morpheme-level segmentation, segmentable morphemes will be separated by hyphens according to the second rule of the LGR.

The LGR also contains a proposed list of category labels which were applied for our corpus. Since this list does not cover all morphological phenomena existing in the examined Uralic languages, two further tagsets were used to supplement the original list of the LGR. One of them is the Wikipedia page of the list of glossing abbreviations[5] which provides a more detailed list of grammatical terms and their abbreviations for interlinear glossing. The other source was the collection of abbreviation lists provided by the reference grammars of the languages in question. When choosing our tags, we preferred the standard category terms and labels over the particular ones.

We follow international standards even in the file formats. All text files are UTF-8 encoded plain text files. The token-level annotations are represented in separate columns of `tsv` files, which can be easily converted into `XML` files or can be directly imported into ELAN as tiers. For transcribing and archiving audio and video data, we use ELAN which is one of the most widely used multimedia annotation tool[6]. ELAN allows audio and video recordings to be time-aligned with multilayered annotations, called tiers. The linguistic annotations and the metadata are stored in a hierarchical structure, which can be exported as standard `XML` files, which can be used as an input of further text processing steps.

[5] https://en.wikipedia.org/wiki/List_of_glossing_abbreviations

[6] http://tla.mpi.nl/tools/tla-tools/elan/

2.3 Consistency

Consistency is the third principal criterium which would apply on all levels: a unified character table has been created to be able to ask one query to the whole corpus; and similarly, the same morphological label is used for the same morphological phenomenon.

Research in language documentation aims at creating long lasting, multipurpose and multifaceted databases in which language technology can definitely help to create systematically annotated corpora, rather than eclectic data collections. The main difference between the former and the latter one is consistency.

Consistency is a basic requirement so that one can ask a query on the whole corpus. One of the great advantages of corpora is that they provide not only separate examples but all instances of the searched term, so analyses based on frequency become available. This important property of corpora can be ensured only if one follows the principle of consistency, and always uses the same appropriate character for representing the same letter and different characters for representing different letters. Therefore, we created a unified character table which contains all possibly used characters in all writing, transliteration and transcription systems of all languages concerned.

As for the level of morphology, the glosses are converted from the output of several morphological analyzers, thus a mapping of the different tags into the set of the unified labels has to be conducted. As a consequence, one morphological phenomenon is always signed with one and the same label.

2.4 Open access

During the whole corpus building process, we aim for following the philosophy of open access, which has two aspects. First, we prefer to use freely available language processing tools and to re-use already collected data sets if it is possible. Second, the results of the project (text and processing resources) will also be freely available.

We plan to make the database available, not only as a downloadable version, but also via an online search interface, which offers the user several features that greatly facilitate the linguistic analysis of large amounts of authentic linguistic data. Moreover, considering the need for long-term preservation in order to assure that these data will be available for future generations, we want to provide the structured data for an international language archive which offers archiving service, such as the Documentation of Endangered Languages (DOBES) corpus hosted by The Language Archive.

3 Text collection and sampling

As mentioned before, the goal of the text selection is to design a corpus that contains reliable, natural, and representative data. There are many factors, however, that one has to consider with respect to the selected Uralic languages when collecting text material. The two main factors discussed here are (i) the level of endangerment of the languages and (ii) the difficulties of data sampling.

Both the vitality and the documentational status vary considerably between the languages of our project. The variation is remarkable especially between the Udmurt language and the languages spoken in Siberia, i.e. Synya Khanty, Surgut Khanty and Tundra Nenets.

The EGIDS level[7] of Udmurt is 5, i.e. it is *developing*, which means that there is literature which is available in a standardized form, though it is not yet widespread or sustainable. In the case of the other three Siberian languages, this value is 6b, i.e. *threatened*, which means that the use of the languages is restricted to the domains of home and family interactions. Native speakers typically belong to the older generation, while the younger ones are losing their heritage languages. Despite the fact that the speaking communities of the three languages in Siberia have territorial autonomies, they are not considered as official languages in the Russian Federation. On the contrary, Udmurt is one of the official languages of Udmurtia. As the languages spoken in Siberia are considered to be of low prestige, efforts of language planning and language revitalization have only sporadically be made in the area. Similarly to several indigenous languages found in Siberia, they are regarded as poorly described and documented languages.

The difference between the levels of vitality results in the limitation or lack of availability of certain text types in the case of the languages spoken in Siberia. While there is a relatively large amount of printed text material collected from speakers who can be characterized as being "old, fluent speakers" of the community (cf. [5]), there is no data from the informal text genres such as blogs or tweets produced by the younger generation. Although several fieldtrips have been undertaken to the traditional territories recently, the collected texts are usually published only in a printed form, or if electronically, they are not available for the research community. If there are electronic corpora or text collections available, they do not provide a representative sample of the language.

Consequently, in the case of the old texts, we selected those folklore texts that were collected in the beginning of the 20th century and (mainly) edited by the fieldworkers themselves. The old Synya Khanty texts come from the collection of Wolfgang Steinitz

[7] https://www.ethnologue.com/about/language-status

[6], which was published in 1975 but collected in the 30s. The old Surgut Khanty texts were collected by Heikki Paasonen [7] in 1900–01 at the Yugan river. As for Udmurt, old text material comes from two sources: once, from the collection of Bernát Munkácsi [8] from 1887, and second, from the collection of Yrjö Wichmann [9] which was published in 1901. The old Tundra Nenets texts are also folklore texts, collected by Toivo Lehtisalo [10] in 1911–12. Even though the genre of these text samples is the same, we tried to keep the ratio of the (sub)dialects, the age and the gender of the informants as balanced as possible. The full table with metadata is available on the web site of the project (the URL will only be provided in the camera ready version).

The sources of the new text material are more diverse. In the case of these texts, we aim to collect and process texts from genres that may potentially represent Russian contact. For this purpose, blogs, interviews appeared in newspapers or in books, and narratives of personal stories have been found suitable. The new Khanty data contains transcribed interviews recently collected during fieldwork. The new Udmurt texts were sampled from the blogs called *Мынам малпанъёсы*[8] and *Марайко*[9]. The new Tundra Nenets data contains interviews from the newspaper entitled *Нярьяна Нгэрм* ('Red North') published in Salekhard. Besides, we acquired and preprocessed several recently collected folklore text samples from the collection of Labanauskas [11] and Pushkareva–Khomich [12]. For each type of new text/data, we clarify their access rights and apply for authorisation.

The new spoken data mainly originate from fieldwork of our project members. These data will be transcribed and time-aligned in ELAN. We plan to collect contemporary data from the same territories from where the old data samples originate, with which we aim at reducing the effects of the influence factors.

4 Text processing

The first step of the corpus building workflow is the acquisition of source data, which typically contains the steps of scanning and OCRing or downloading from the web, see Section 4.1. The language documenters used Latin-based transcription systems, while the languages concerned use Cyrillic writing systems, and we aim at creating the IPA transliteration for the text samples in all languages, therefore we have to deal with several transcriptions and transliterations, see for details in Section 4.2. The morphological analysis and disambiguation is discussed in Section 4.3.

[8] `http://udmurto4ka.blogspot.hu/`

[9] `http://marjamoll.blogspot.hu/`

4.1 Acquisition of the original text material

A significant part of the linguistic material was only available in print. In this case, digitization was carried out by scanning followed by a conversion process from the scanned images into regular text files aided by an OCR software.

Since we work with several writing, transcription and transliteration systems, a key aspect of an OCR software was its ability to be trained. For this purpose, we used the Abbyy FineReader Professional edition[10], which can be trained in an interactive way and produces a fairly good quality result.

Some new text samples were acquired by downloading them from the web. In these cases, the plain text had to be extracted from `HTML` sources or `PDF` files. Because of the diverse sources, a kind of character-level normalization is needed. For this purpose, a `Perl` script is used which lists all of the Unicode characters used in the document. Based on this, foreign language parts and not properly used characters can be removed or changed, such as in the case of the Cyrillic letter el with descender, see Section 2.2.

4.2 Transcription and transliteration

The database contains each text material at least in its original transcription used by the language documenter and in IPA transliteration. Moreover, since the writing system of the languages concerned is based on the Cyrillic alphabet, we preserve the original Cyrillic script, if it is available. If not, we create it in a conversion step, if it is needed for the morphological analysis. Since some morphological analyzers only accepts the input in a given FUT transcription, texts in some languages had to be converted into that transcription as well.

The Finno-Ugric transcription (FUT) system or the Uralic Phonetic Alphabet is a phonetic transcription system first published by Eemil Nestor Setälä [13] and only used in the field of Uralistics. It is called one system, however, it is actually a common name for several transcription systems developed and used by researchers and language documenters in the Uralic studies from the end of the 19th century until recently.

The systems of Wolfgang Steinitz, Márta Csepregi [14], Bernát Munkácsi, Yrjö Wichmann, Toivo Lehtisalo and Péter Hajdú [15] differ from each other very much. Each of them follow his/her own inner logic, which is sometimes hard to be detected, additionally they are not consequent in the sense that they frequently apply different characters for the same sound and vice versa. For example, in the case of

[10] `http://finereader.abbyy.com/`

Surgut Khanty, the close central rounded vowel represented with the IPA symbol /ʉ/ is marked with /ü/ by Steinitz and with /ŭ/ by Márta Csepregi.

There are altogether 11 conversion directions for the four languages, as detailed below. The old Synya Khanty texts are originally transcribed by Steinitz, who used his own FUT-like system, which has been converted first into IPA, and second into another FUT-like transcription used by the developers of the morphological analyzer for Synya Khanty (see later in Section 4.3).

The old Surgut Khanty texts were kindly provided us by Elena Skribnik and Zsófia Schön, members of the Ob-Ugric Database (OUDB) research project[11]. The OUDB text corpus contains the texts only in IPA transliteration, thus the Paasonen texts are available only in IPA. However, the modern Surgut Khanty texts are written with Cyrillic characters, which have been converted first into the transcription system of Márta Csepregi, then from that into IPA.

As for the Udmurt language, we had to create conversion rules for four directions. Once, we compiled the conversion rules from the transcription of Bernát Munkácsi and Yrjö Wichmann into IPA. Since the available morphological analyzers for Udmurt (see Section 4.3) only accept Cyrillic script, we had to create the transliteration rules for the direction from IPA to Cyrillic. In the case of modern Udmurt texts, the inverse direction is used, since we have to create the IPA transliteration of texts written in Cyrillic as well.

The old Tundra Nenets texts transcribed by Lehtisalo are exceptions in the sense, that they were OCRed directly in the transcription system of Péter Hajdú, not in that of Lehtisalo. The reason behind this is that Lehtisalo's transcription is unduly difficult, and a part of the characters could not be represented by standard Unicode characters. Therefore, the starting point of the conversion was the Hajdú transcription which was converted into IPA and then into Cyrillic, the latter one for being the input of the morphological analyzer. And last but not least, the modern Cyrillic Tundra Nenets texts are also converted into IPA.

First, manually compiled transcription rules were created for all directions by linguist experts of the languages concerned. These rules were then transformed into substitution commands expressed by extended regular expressions which can be fed to the Unix command `sed` with the option `-f`. Thus, it is a typical rule-based system, one of whose shortcomings is that it is language-dependent, i.e. the system is not portable to other languages or to other directions without changing. Since the rules must be ordered, it can be quite difficult to incorporate new rules into the system. Moreover, it is hard to keep track of all rules and a single error may cause the system to malfunction. However, rule-based systems have advantages as well, namely

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

the high precision they achieve. Since all of the converted texts are checked and corrected by linguist experts, we rather vote for higher precision, even at the cost of difficulty.

During the process of the transcriptions one barrier arose which is resulted from the lack of the standard spelling of the old and modern texts. In both cases, we avoided to standardize the different forms that appeared in the texts even when they apparently represented the same word. We decided to keep these "inconsistent" forms, because the standardizing process may easily lead to the loss of important information regarding the phonetic system of the languages.

4.3 Morphological analysis

The corpus will contain rich linguistic annotation even on the morphological level. For each token, its lemma, its POS tag and its English gloss will be added as annotations. As mentioned in Section 2.2, the morphological annotations will be carried out by the application of the available morphological analyzers, by the conversion of glosses from their output, and by the manual correction of the output of the conversion. Therefore, if the morphological analyzer is capable to create morpheme-by-morpheme segmentation, the tokens will also be morph-level segmented.

There are existing morphological analyzers for three of the four languages concerned, with which the process of the morphological annotation can be supported, however the whole annotation process cannot be conducted fully automatically.

The most well-known text processing framework for under-resourced Uralic languages is Giellatekno[12]. It provides a fully established framework for creating language processing tools, such as proofing tools, digital dictionaries and morphological analyzers. The latter one has been developed for Udmurt, Northern Khanty and Tundra Nenets.

Besides, there are morphological tools developed for small Uralic languages, such as Udmurt and Synya Khanty, by a Hungarian language technology company (MorphoLogic) and the Research Institute for Linguistics of the Hungarian Academy of Sciences [16]. These analyzers are not open source tools, but they are available via an online interface[13]. They output an `HTML` file containing all potential analyses of each token. To support the process of manual checking and disambiguation, we use a web-based interface which was originally created for the disambiguation of Old Hungarian texts [17]. The proper analysis can be chosen from a pop-up menu containing a list of possible analyses which appears when the mouse cursor is placed over the

[12] `http://giellatekno.uit.no/`

[13] `http://www.morphologic.hu/urali/`

word. The analyses provided by the Giellatekno analyzers are also converted so that to be fed to this web-based interface.

For the Synya Khanty texts, we use the analyzer of MorphoLogic. Its reason is that the Giellatekno analyzer needs Cyrillic input, while all of our sources are in some Latin-based transcriptions.

As far as we know, the Surgut Khanty language is the only one of the four languages we deal with for which there is no morphological analyzer. Based on Zipf's law, the large part of the text is covered by the most frequent words, thus if we provide the gloss for the first n most frequent words from a lookup, the pains of the manual work can be highly reduced. For this reason, a linguist expert created a table with the morphological codes and the English translation of the lemma for the 122 most frequent words of the modern texts, from which the glosses for the 64% of the text can be automatically generated.

For the morphological analysis of the Udmurt texts, we use the analyzer of MorphoLogic, since it provides morph-level segmentation and Hungarian translation as well.

In the case of Tundra Nenets, we use the Giellatekno analyzer. However, the grammar files in Giellatekno describe another dialect and follow another grammar. For this reason, we plan to create new grammar files for Tundra Nenets within the framework of Giellatekno, which may be used for morphological analysis of this dialect of Tundra Nenets according to the grammar of Nikolaeva [18] in the future.

5 The structure of the corpus

The corpus has three main annotation levels, each of which has an obligatory version which must be created for each text sample. Two annotation levels are token-level aligned, while one is sentence-level aligned. The first level is the level of the original text itself, which can be written either in Cyrillic or in a FUT transcription, but it must have an IPA transliteration, which is the obligatory version at this level. Token-level alignment means in this case that several transcriptions and transliterations of each token can be seen side-by-side in the `tsv` files. The morphological annotation for each token contains the lemma, the POS tag and the English gloss. The third annotation level is the level of translations, where the obligatory translation is the English one, however there are several text samples which have German, Hungarian or Russian translation as well. The different language translations are sentence-level aligned.

The token- and sentence-level aligned text samples will be imported into ELAN, where spoken data will also be sentence-level time-aligned. ELAN presents the annotation levels as horizontal tiers, which is illustrated by the Tundra Nenets example

YRK Hajdú:	jā	mīdaxana	amkerta	jaŋkūwi
YRK IPA:	ja	mi:daxana	ămkerta	jăŋkuwi
YRK Cyrillic:	я	мыдахана	амкэрта	яӈкувы
lemma:	я	мы	ӈамгэ	яӈгось
POS:	N	Ptcp	Pron.neg	V
gloss:	earth	create.IPFV.PTCP.LOC	something.CONC	neg.EX.INFER.3SG
ENG:	when the earth was created, there was nothing			
GER:	zur zeit der erschaffung der erde gab es nichts			
HUN:	a Föld teremtésének idején nem volt semmi			

Table 1: Token- and sentence-level aligned text sample in Tundra Nenets.

in Table 1.

6 Conclusion and future work

This paper discussed the most important theoretical considerations and the process of building a linguistically annotated database of certain endangered Uralic languages within the framework of the research project called *Languages under the Influence*. The general theoretical considerations applied in our work cover the main principles of language documentation, the use of international standards, the principle of consistency and open access. We addressed problems of collecting text samples in endangered languages and provided solution to resolve these specific problems.

Our database contain token- and sentence-level aligned data for four languages: Tundra Nenets, Udmurt, Synya Khanty and Surgut Khanty. Each text sample is available in at least in IPA transliteration, extended with morphological information and with English translation. The still under construction thus ever growing database is freely available via the URL of our web site which will be provided in the camera ready version of the paper.

Since it is an ongoing project, several future directions emerge during each corpus building step. We aim for creating a fully annotated database containing at least 4000 tokens text samples from the two time period for all the aforementioned languages until the end of the pilot project. Long-term future plans highly depend on the result of the ERC project proposal.

As mentioned, all results of the pilot project will be freely available. Not only the converted, translated and morphologically annotated text samples are freely available, but we will provide all conversion rules and tools as well as the table of the unified morphological tagset with the unified character table. We also plan to develop an online search interface which offers several features.

Acknowledgments

The research presented in this article has been funded by the Hungarian National Research, Development and Innovation Office (grant ID: ERC_HU_15 118079).

Database building is a group effort; beside the authors, the following researchers have been contributing to it: Erika Asztalos, Katalin Gugán, Ágnes Kalivoda, Nóra Lien Nguyen-Dang, Eszter Ruttkay-Miklián, Orsolya Tánczos.

Our thanks go to: Zsófia Schön who helped us in the conversion of modern Surgut Khanty texts; Elena Skribnik and Zsófia Schön for kindly providing us the glossed texts of Paasonen; A. S. Pesikova and A. N. Volkova who gave us permission to use their interview transcriptions.

References

[1] Peter K. Austin. Language documentation in the 21st century. *JournaLIPP*, (3):57–71, 2014.

[2] Anthony C. Woodbury. Language documentation. In Julia Austin, Peter K.; Sallabank, editor, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press, 2011.

[3] Nikolaus P. Himmelmann. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation*, 6:187–207, 2012.

[4] Rogier Blokland, Marina Fedina, Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. Language documentation meets language technology. In *First International Workshop on Computational Linguistics for Uralic Languages*, number 2 in Septentrio Conference Series, pages 8–18, 2015.

[5] Colette Grinevald and Michel Bert. Speakers and communities. In Julia Austin, Peter K.; Sallabank, editor, *The Cambridge Handbook of Endangered Languages*, pages 45–65. Cambridge University Press, 2011.

[6] Wolfgang Steinitz. *Ostjakologische Arbeiten*. Akadémiai Kiadó, Budapest, 1975.

[7] Edith Vértes, editor. *Heikki Paasonens surgutostjakische Textsammlungen am Jugan. Neu transkribiert, bearbeitet, übersetzt und mit Kommentaren versehen von Edith Vértes*, volume 240 of *Mémoires de la Société Finno-Ougrienne.* Suomalais-Ugrilainen Seura, Helsinki, 2001.

[8] Bernát Munkácsi. *Votják népköltészeti hagyományok.* Magyar Tudományos Akadémia, Budapest, 1887.

[9] Yrjö Wichmann. *Wotjakische Sprachproben II. Sprichwörter, Rätsel, Märchen, Sagen und Erzählungen.* Helsinki, 1901.

[10] Toivo Lehtisalo. *Juraksamojedische Volksdichtung.* Suomalais-Ugrilainen Seura, Helsinki, 1947.

[11] К. И. Лабанаускас. *Ненецкий фольклор. Мифы, сказки, исторические предания. Вып. 5.* Красноярск, 1995.

[12] Е. Т. Пушкарёва and Л. В. Хомич. *Фольклор ненцев.* Новосибирск, 2001.

[13] Eemil Nestor Setälä. Über transskription der finnisch-ugrischen sprachen. *Finnisch-ugrische Forschungen*, 1:15–52, 1901.

[14] Márta Csepregi. *Szurguti osztják chrestomathia.* Szeged, 2011.

[15] Péter Hajdú. *Chrestomathia Samoiedica.* Tankönyvkiadó, Budapest, 1989.

[16] Attila Novák. Morphological Tools for Six Small Uralic Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 925–930. ELRA.

[17] Attila Novák, György Orosz, and Nóra Wenszky. Morphological annotation of Old and Middle Hungarian corpora. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 43–48, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[18] Irina Nikolaeva. *A Grammar of Tundra Nenets.* Mouton de Gruyter, 2014.