Language technology resources and tools for Mansi: an overview

Csilla Horváth[1], Norbert Szilágyi[2], Veronika Vincze[3], Ágoston Nagy[2]
[1]University of Szeged, Institute of English-American Studies
cshorvath@ieas-szeged.hu,nagyagoston@yahoo.com
[2]University of Szeged, Department of Finno-Ugric Studies
norbertszilagyi91@gmail.com
[3]University of Szeged, Department of Informatics
vinczev@inf.u-szeged.hu

Abstract

In this paper, we offer an overview of language technology tools and resources (being) developed for an endangered minority language, Mansi. We pay special attention to lexical resources and morphological analyzers. Online dictionaries, morphological analyzers and a corpus are already available (or will be made available soon) for the language, which are described in the paper. Moreover, we also briefly present our efforts to contribute to the field of Mansi language technology. In several cases the weaknesses of existing resources or tools motivated us to implement some new tools, which are also presented in the paper. All of the tools and resources developed by us will be made freely available for the research community and anyone interested.

1 Introduction

According to a UNESCO review (UNESCO 2003) 50-90% of the known languages on Earth are likely to become extinct by the end of the century, thus any research on the unobserved aspects of endangered languages requires no further justification. Besides the importance of language documentation, the necessity of sociolinguistic or

The 3rd International Workshop for Computational Linguistics of Uralic Languages, pages 56–65,
St. Petersburg, Russia, 23–24 January 2017.

anthropological linguistic research as well as creating tools for language users and language learners is unquestionable.

The urbanization of indigenous people of Siberia has rarely been studied from a linguistic, ethnographic or cultural anthropological viewpoint, which calls for immediate reaction as excluding cities from fieldwork may often result in neglecting the majority of the Siberian population. This is especially true for people belonging to minorities in Russia: as for Obi-Ugric people, about a half of them cannot participate in linguistic and ethnographic studies [1]. Hence, sociolinguistic research of the Mansi language, especially revitalization efforts and the usage of urban language is of utmost importance.

In this paper, we aim at offering a landscape of language technology resources and tools for an endangered minority language, Mansi. Mansi (former term: Vogul) is an extremely endangered indigenous Uralic (more precisely Finno-Ugric, Ugric, Ob-Ugric) language, in minority position, spoken in Western Siberia, especially on the territory of the Khanty-Mansi Autonomous Okrug. Among the approximately 13,000 people who declared to be ethnic Mansi according to the data of the latest Russian federal census in 2010 only 938 stated that they could speak the Mansi language, which shows drastic decrease compared with the 2002 census figures (2,746 speakers).

The Mansi have been traditionally living on hunting, fishing, to a lesser extent also on reindeer breeding, they got acquainted with agriculture and urban lifestyle basically during the Soviet period. Urban lifestyle gained prominent importance in the last decades, the proportion of Mansis living in large villages, towns and cities has been constantly growing and by 2010 exceeded half of the Mansi population (57%). This tendency together with the unsatisfactory access to Mansi education and the weakening intergenerational transmission of the language makes the situation of Mansi vulnerable. Besides their scientific and academic novelty, the vulnerability of Mansi language also underlines the importance of creating computational tools for Mansi, for the benefit of scholars and researchers as well as language learners and language users.

The principles of Soviet linguistic policy according to which the Mansi literary language has been designed kept changing from time to time. After using Latin transcription for a short period, Mansi language planners had to switch to the Cyrillic transcription in 1937. While until the 1950s the more general tendency was to create new Mansi words to describe the formerly unknown phenomena, later on the usage of Russian loanwords became more dominant. Since the 1990s the tendencies governing the planning of Mansi language use and language acquisition have become multidimensional, important differences and interferences may be observed between the different actors of language use, especially the leading specialists (mainly following the Soviet academic policy) and the journalists (using and promoting the language

on a daily basis, with the largest active number of followers). During our work we try to take every approach into consideration, when it is impossible to compromise (e.g. questions related to orthography), we tend to follow the examples found in majority language use.

In this paper, we aim at summarizing the current situation of Mansi within computational linguistics: we give a summary of the existing tools and resources (with special regard to lexical resources and morphological analyzers) and we will also briefly present our efforts to contribute to the field.

2 Lexical resources

2.1 Online dictionaries for Mansi

There are only a handful of dictionaries available for Mansi. The vocabularies collected by the early researchers of Mansi, Munkácsi's and Kannisto's (published as late as [2] and [3]) serve as great contribution and unfailing source of data for researchers. Nonetheless, they are inadequate for creating computational tools that would be equally useful for researchers and Mansi language users as well, partly because both dictionaries use especially detailed Latin-based script with plenty of diacritics, and partly because both dictionaries incorporate materials of different dialects and subdialects that have become extinct since the days of Munkácsi's and Kannisto's fieldworks. We are also aware of one small dictionary for Mansi (available at `http://glosbe.com/mns`), though this resource is limited in size as it is based on about 200 translated sentences, moreover, it applies Cyrillic and Latin transcriptions inconsistently.

Thus, instead of starting from the larger dictionaries compiled by European researchers, we decided to begin our work with the smaller, but more actively used dictionaries published by Mansi researchers, that is the Mansi-Russian-Mansi dictionary by Rombandeeva and Kuzakova (4,000 entries) [4] and the Russian-Mansi dictionary by Rombandeeva (11,000 entries) [5]. These dictionaries use Cyrillic script, contain most of the currently used Mansi vocabulary, and are often consulted by native Mansi or specialists working with the Mansi language and they are widely used in Mansi education.

The process of dictionary building is the following: the automatic optical character recognition is followed by manual correction and translation of the entries, and then this database is turned into a searchable, digitized dictionary [6], as detailed below.

The beta version of the online Mansi dictionary now contains approximately 20,000

entries[1]. The Mansi forms were retrieved from the PDF versions of Rombandeeva's and Kuzakova's [4], as well as Rombandeeva's [5] dictionaries by means of optical character recognition, then lexical entries from different sources were merged. It was also noted which dictionary and which page they come from. Translations that included several synonyms were added as separate items. The Mansi lexemes are supplemented with the Russian translation given by the dictionaries, and Hungarian (complete) and English (in progress) translations provided by linguists of the FinugRevita research group, parts of speech and annotation of the sources, i.e. the dictionaries that are contained within. In addition, information about their morphological paradigm is also encoded, as well as the argument structure of certain verbs.

The online Mansi dictionary being a key resource for creating a morphological analyzer, the project also aims to make it available for public use as well, thus meeting a long-felt need for a sufficient Mansi–English–Mansi and a suitable online Mansi dictionary. After some technical refinements, our online Mansi dictionary will be made available on the Giellatekno website (`http://giellatekno.uit.no`).

We also plan to collate our dictionary with the data on the Northern Mansi dialect group of Munkácsi's enormous Mansi–Hungarian dictionary [2] – by relying on its simplified transcipt by Béla K'almán – and also expand it with the Northern Mansi material of Balandin's and Vakhrusheva's Mansi–Russian dictionary [7], as well as with dozens of the most necessary neologisms describing different features of contemporary lifestyle (such as the urban environment, oil mining or judicial terms), created and used first and foremost by the journalists of the Mansi newspaper *Luima Seripos* (see below).

2.2 Mansi Wordnet

Recently, another lexical resource has been constructed for Mansi: [8] report on the construction of a wordnet for Mansi. Special challenges were met during the building process, among which the most important ones are the low number of native speakers, the lack of thesauri and the bear language. The bear is a prominently sacred animal venerated by Mansi, thus triggering a detailed taboo language. Since the bear is thought to understand the human speech, it is required to use taboo words while speaking about the bear, the parts of its body, or any activity connected with the bear (especially bear hunting) so that the bear would not understand it. Thus vocabulary items belonging to the bear language had to be separately indicated in the Mansi wordnet, which currently contains about 300 synsets and is under constant

[1]Note that the number of the entries exceeds the sum of the entries of the individual ditionaries. This is due to synonyms and different senses of Mansi words, which were grouped under the same head entry in the Russian–Mansi dictionary, but are now counted separately.

development. As for the proportion of part-of-speech categories, nouns prevail over verbs with 210 nouns (70%) and 90 verbs.

3 Mansi Morphological Analyzers

Mansi is a morphologically rich language, similar to other Uralic languages. Thus, its automatic morphological processing requires a properly designed morphological analyzer. For Mansi, there already exists a morphological analyzer [9] developed by MorphoLogic Ltd.[2]. However, this has several issues that are problematic concerning contemporary Mansi language use. First, it employs Latin-based transcription used by scholars but the Mansi orthography used by the speakers themselves is Cyrillic-based. Second, its vocabulary is based on Munkácsi's Mansi dictionary [2] and it was optimized for the texts covered in Kálmán's *Chrestomathia Vogulica* [10] and *Wogulische Texte* [11], mostly collected at the end of the 19th and the first half of the 20th century. Hence, the contemporary lexicon and genres of the 20th and 21st centuries are underrepresented. Third, it is not open-source.

For all the above mentioned reasons, we chose to create a new morphological analyzer for Mansi from scratch. From among the many currently available finite-state tools, the HFST standard was chosen in order that the analyzer could be integrated into the framework which is used at the GiellaTekno website. This choice is thus motivated by the fact that in this way, the morphological analyzer can be integrated into a large system dealing with minority languages with a common interface. The files in the morphological analyzer can be grouped into two categories: stems and affixes. Mansi words (stems) are given in a lexicon, together with morphological information and their Russian translations, in addition, there are morphological rules that are responsible for analyzing and generating different inflectional forms of the individual stems. Furthermore, stems and affixes are organized into different files on the basis of their part-of-speech category, since nouns are conjugated differently than verbs, and the whole system is easier to modify and to look through.

The dictionary mentioned in Section 2.1 serves as a basis for the lexicon integrated into the morphological analyzer. Nevertheless, the original paper-based dictionaries were published decades ago, which means that the lexicon cannot contain contemporary terms of vocabulary (e.g. those related to internet and social media). Thus, it is constantly expanded by adding novel lexical items in a semi-automatic way: the analyzer is regularly run on contemporary texts and if some of the words cannot get a proper morphological analysis, their stem is added to the lexicon.

[2] http://www.morphologic.hu/urali/index.php?lang=hungarian&a_lang=chv

Another issue with stems is that multiword items were extremely frequent in the dictionaries. For example, the учитель 'schoolteacher $masc.$' could be translated as нӈврамыт ханисьтан хум built up of the element *children-teaching man* , and the feminine counterpart учительница 'schoolteacher $fem.$' as нӈврамыт ханисьтан нэ̄ from *children-teaching woman*. However, the HFST formalism could not support multiword items and they needed to be reduced to their syntactic head, in this case, хум and нэ̄.

As for the morphological rules, lexical entries of Mansi were grouped into different morphological categories depending on the inflectional paradigm they belong to. For this, we relied on the descriptions found in several Mansi grammars [12, 13], as well as on the linguistic intuitions of native speakers of Mansi.

In order to classify the Mansi stems into inflectional paradigms, we analyzed the words' phonological structure. First, we listed all the possible syllables that can occur at the beginning of the word, at the end of the word, or within words. This was an important step in establishing inflectional paradigms as in the case of nouns, it is the last (two) syllable(s) that determine the quality of affixes and linking vowels, whereas in the case of verbs, it is the number of syllables and the quality of the last syllable. Then, inflectional paradigms were created for each type separately in order to be able to analyze and generate all inflectional forms of Mansi nouns and verbs. Right now, the system includes 36 nominal and 27 verbal paradigms.

Here we offer a short sample of the nominal stems and inflectional paradigms for illustrative purposes. The first column denotes the stem and its inflectional stem. The second column contains the inflectional paradigm it belongs to, e.g. N_CVS_masnut__n means that it is a nominal paradigm for nouns that end in a combination of a vowel and a softening consonant, a typical example of which is *masnut* 'cloth'.

```
миркол:миркол   N_CVS_masnut__n   "сельсовет" ;
щӗмья:щӗмья     N_VO_maa__n       "семья" ;
нэ̄пак:нэ̄пак    N_CVH_luw__n      "книга" ;
ӯльпа:ӯльпа     N_VO_maa__n       "кедр" ;

LEXICON N_VO_maa__n
N_VO_maa__n ;

! non possessive forms

+N+Sg+Nom: K ;
+N+Sg+Loc:т K ;
+N+Sg+Lat:н K ;
```

+N+Sg+Abl:ныл K ;
+N+Sg+Ins:л K ;
+N+Sg+Tra:г K ;

+N+Du+Nom:г K ;
+N+Du+Loc:гт K ;
+N+Du+Lat:гн K ;
+N+Du+Abl:гныл K ;
+N+Du+Ins:гтыл K ;

+N+Pl+Nom:т K ;
+N+Pl+Loc:тт K ;
+N+Pl+Lat:тн K ;
+N+Pl+Abl:тныл K ;
+N+Pl+Ins:тыл K ;

Based on the above information, the following morphological analyses are provided by the system for the word щёмьят “families” or “in family”:

щёмьят щёмья+N+Pl+Nom
щёмьят щёмья+N+Sg+Loc

Beside parts-of-speech that can be inflected, the morphological analyzer also includes words belonging to parts-of-speech that cannot be inflected. For instance, adverbs, conjunctions and interjections are also adequately recognized by the system.

Our Mansi morphological analyzer will be made freely available soon within the Giellatekno infrastructure.

4 Mansi corpora

In order to test our morphological tools, we have started to create a Mansi corpus, which consists of the articles published in the Mansi newspaper *Luima Seripos* (Mansi for “Northern dawn”). The online archive of *Luima Seripos* is available on the homepage of the joint editorial board of *Luima Seripos* and regional Khanty newspaper *Khanty Yasang*.[3]. The Mansi texts published in *Luima Seripos* cover various topics

[3] http://www.khanty-yasang.ru/luima-seripos/archive

such as traditional lifestyle, folklore and short biographies, as well as domains of urban life.

Currently, the corpus contains issues of *Luima Seripos* from 2013 (more precisely, issues 1050-1131). The corpus consists of 520,000 tokens, converted into XML format. Cyrillic characters with diacritics, i.e. those that denote vowel length were segmented into two characters: the vowel itself and the macron, which enables a proper display of the Unicode characters (the original website is sometimes inadequaly displayed on certain machines or browsers). We have been constantly working on the extension of the corpus and we are planning to add about 150,000 tokens to it.

The following metadata are also assigned to the texts of the corpus: number of issue, date of publication, author of the article, title of the article, link to the article and a unique identifier for each article within the corpus. Also, the XML file applies special tags for named entities such as person names and locations as well as embedded texts written in Russian.

As work in progress, we would like to mention that about 5,000 tokens of the corpus are being manually annotated for part-of-speech tags and syntactic structures so that later on we can test and evaluate our morphological analyzer and POS tagger under development.

5 Conclusions

In this paper, we offered an overview of language technology tools and resources (being) developed for Mansi. Online dictionaries, morphological analyzers and a corpus are already available (or will be made available soon) for the language, however, there is still room for improvement. In addition to the constant update and extension of the above mentioned tools, we intend to create a small dependency treebank of Mansi in harmony with the Universal Dependencies project [14], which may enhance the implementation of a syntactic parser for Mansi. To reach the widest possible range of active users we consider the opportunity of creating online games based on the already existing computational tools. For the Mansi language users and language learners we plan to produce online games activating the users knowledge on Mansi vocabulary. We intend to make all of our resources and tools freely available to anyone interested.

The majority of the future Mansi audience of computational language tools, such as children, teenagers and young adults were raised in multiethnic families and multiculticultural settlements, most of the cases in towns and cities. Only 0.72% of the total population of the Khanty–Mansi Autonomous Okrug belongs to Mansi ethnicity, this proportion occasionally may be higher in cities, as for example in the capital of the Okrug, Khanty–Mansiysk (1.5%). This condition together with the weakening

intergenerational transmission of the language and the just gradually changing educational system make the situation of Mansi vulnerable. The young Mansi language users and language learners are familiar with the newest technological developments, in general they have internet access and use the social media as well, most of them got acquainted with computational language teaching tools as well in alternative educational institutions [15]. Thus, besides their scientific and academic novelty, the creation of computational tools for the Mansi language is crucial for the benefit of the language learners and language users as well.

Acknowledgments

This work was supported in part by the Finnish Academy of Sciences and the Hungarian National Research Fund, within the framework of the project *Computational tools for the revitalization of endangered Finno-Ugric minority languages (FinUgRevita).* Project number: OTKA FNN 107883; AKA 267097.

References

[1] Z. Nagy. Szibéria néprajza és a város: Akik kimaradtak az összefoglalókból. In S. Szeverényi and T. Szécsényi, editors, *Érdekes nyelvészet*, 2016.

[2] B. Munkácsi and B. Kálmán. *Wogulisches Wörterbuch.* Akadémiai Kiadó, Budapest, 1986.

[3] Artturi Kannisto. *Wogulisches Wörterbuch.* Helsinki, 2013. Kotimaisten Kielten Keskuksen Julkaisuja 173.

[4] Е. И. Ромбандеева and Е. А. Кузакова. *Словарь мансийско-русский и русско-мансийский.* Просвещение, Ленинград, 1982.

[5] Е. И. Ромбандеева. *Русско-мансийский словарт.* Миралл, Санкт-Петербург, 2005.

[6] N. Thieberger and A. L. Berez. Linguistic Data Management. In N. Thieberger, editor, *The Oxford Handbook of Linguistic Fieldwork*, chapter 4, pages 90–118. Oxford University Press, Oxford, 2012.

[7] А. Н. Баландин and М. П. Вахрушева. *Мансийско-русский словарь с лексическими паралеллями из южно-мансийского (кондинского) диалекта.* Просвещение, Ленинград, 1958.

[8] Csilla Horváth, Ágoston Nagy, Norbert Szilágyi, and Veronika Vincze. Where Bears Have the Eyes of Currant: Towards a Mansi WordNet. In Verginica Barbu Mititelu, Corina Forascu, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Eighth Global WordNet Conference*, pages 130–134, Bucuresti, Romania, 2016.

[9] Gábor Prószéky. Endangered Uralic Languages and Language Technologies. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 1–2, Hissar, Bulgaria, September 2011.

[10] B. Kálmán. *Chrestomathia Vogulica.* Tankönyvkiadó, Budapest, 1963.

[11] Béla Kálmán. *Wogulische Texte mit einem Glossar.* Akadémiai Kiadó, Budapest, 1976.

[12] T. Riese. *Vogul.* Number 158 in Languages of the World/Materials. Lincom Europa, München - New Castle, 2001.

[13] Е. И. Ромбандеева. *Мансийский (вогульский) язык.* Наука, Москва, 1973.

[14] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[15] Csilla Horváth. Old problems and new solution: Teaching methods in the governmental and alternative Mansi educational institutions. *Finnisch-Ugrische Mitteilungen*, 39:37–48, 2013.