# Where Bears Have the Eyes of Currant: Towards a Mansi WordNet

**Csilla Horváth**[1], **Ágoston Nagy**[1], **Norbert Szilágyi**[2], **Veronika Vincze**[3]
[1]University of Szeged, Institute of English–American Studies
Egyetem u. 2., 6720 Szeged, Hungary
horvathcs@ieas-szeged.hu, nagyagoston@lit.u-szeged.hu
[2]University of Szeged, Department of Finno-Ugrian Studies
Egyetem u. 2., 6720 Szeged, Hungary
norbertszilagyi91@gmail.com
[3]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
Tisza Lajos krt. 103., 6720 Szeged, Hungary
vinczev@inf.u-szeged.hu

## Abstract

Here we report the construction of a wordnet for Mansi, an endangered minority language spoken in Russia. We will pay special attention to challenges that we encountered 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. We will discuss our solutions to these issues, which might have some theoretical implications for the methodology of wordnet building in general.

## 1 Introduction

Wordnets are lexical databases that are rendered according to semantic and lexical relations between groups of words. They are supposed to reflect the internal organization of the human mind (Miller et al., 1990). The first wordnet was constructed for English (Miller et al., 1990) and since that time, wordnets have been built for several languages including several European languages, mostly in the framework of EuroWordNet and BalkaNet (Alonge et al., 1998; Tufiş et al., 2004) and other languages such as Arabic, Chinese, Persian, Hindi, Tulu, Dravidian, Tamil, Telugu, Sanskrit, Assamese, Filipino, Gujarati, Nepali, Kurdish, Sinhala (Tanács et al., 2008; Bhattacharyya et al., 2010; Fellbaum and Vossen, 2012; Orav et al., 2014). Synsets within wordnets for different languages are usually linked to each other, so concepts from one language can be easily mapped to those in another language. Wordnets can be beneficial for several natural language processing tasks, be it mono- or multilingual: for instance, in machine translation, information retrieval and so on.

In this paper, we aim at constructing a wordnet for Mansi, an indigenous language spoken in Russia. Mansi is an endangered minority language, with less than 1000 native speakers. Most often, minority languages are not recognized as official languages in their respective countries, where there is an official language (in this case, Russian) and there is one or there are several minority languages (e.g. Mansi, Nenets, Saami etc.). Hence, the speakers of minority languages are bilingual, and usually use the official or majority language in their studies and work, and the language of administration is the majority language as well. However, the minority language is typically restricted to the private sphere, i.e. among family members and friends, and thus it is mostly used in oral communication, with only sporadic examples of writing in the minority language (Vincze et al., 2015). Also, the cultural and ethnographic background of Mansi people may affect language use: certain artifacts used by Mansi people that are unknown to Western cultures have their own vocabulary items in Mansi and vice versa, certain concepts used by Western people are unknown to Mansi people, therefore there are no lexicalized terms for them.

The construction of a Mansi wordnet help us explore how a wordnet can be built for a minority language and also, an endangered language. Thus, we will investigate the following issues in this paper:

- What are the specialties of constructing a wordnet for a minority language?
- What are the specialties of constructing a word-

net for an endangered language?

- What are the specialties of constructing a wordnet for Mansi?

The paper has the following structure. First, the Mansi language will be shortly presented from linguistic, sociolinguistic and language policy perspectives. Then our methods to build the Mansi wordnet will be discussed, with special emphasis on specific challenges as regards endangered and minority languages in general and Mansi in particular. Later, statistical data will be analysed and our results will be discussed in detail. Finally, a summary will conclude the paper.

## 2 The Mansi Language

Mansi (former term: Vogul) is an extremely endangered indigenous Uralic (more precisely Finno-Ugric, Ugric, Ob-Ugric) languages, 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.

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. 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. As a result of these tendencies some of the terms describing contemporary environment, urban lifestyle, the Russian-dominated culture are Russian loanwords, while others are Mansi neologisms created by Mansi linguists and journalists. It is not uncommon to find two or even three different synonyms describing the same phenomena (for example, hospital): by the means of borrowing the word from Russian (больница), or using the Russian loanword in a form adapted to the Mansi phonology (пӯльница), or using a Mansi neologism to describe it (ма̄хум пусмалтан кол, 'a house for healing people, hospital', as opposed to ня̄врам пусмалтан кол 'children hospital, children's clinic' or ӯйхул пусмалтан кол 'veterinary clinic').

## 3 Semi-automatic construction of the Mansi WordNet

In this section, we will present our methods to construct the Mansi WordNet. We will also pay special attention to the most challenging issues concerning wordnet building.

### 3.1 Low number of native speakers

The first and greatest problem we met while creating the Mansi wordnet was that only a handful of native speakers have been trained in linguistics. Thus, we worked with specialists of the Mansi language who have been trained in linguistics and technology, but do not have native competence in Mansi.

As it is not rentable to build a WordNet from scratch and as our annotators are native speakers of Hungarian, we used the Hungarian WordNet (Miháltz et al., 2008) as a starting point. First, we decided to include basic synsets, and the number of the synsets is planned to be expanded continuously later on. We used Basic Concepts – already introduced in EuroWordNet – as a starting point: this set of synsets contains the synsets that are considered the most basic conceptual units universally.

### 3.2 Already existing resources

In order to accelerate the whole task and to ease the work of Mansi language experts, the WordNet creating process was carried out semi-automatically. Since there is no native speaker available who could solve the problems requiring native competence, we were forced to utilize the available sources as creatively as possible.

First, the basic concept sets of the Hungarian WordNet XML file were extracted and at the same time, the non-lexicalized elements were filtered as in this phase, we intend to focus only on lexicalized elements.

Second, we used a Hungarian-Mansi dictionary to create possible translations for the members of the synsets. The dictionary we use in the process is based on different Mansi-Russian dictionaries (e.g. Rombandeeva (2005), Balandin and Vahruševa (1958), Rombandeeva and Kuzakova (1982)). The translation of all Mansi entries to Hungarian and to English in the new dictionary is being done independently of WordNet developing (Vincze et al., 2015).

In order not to get all Hungarian entries of the WordNet translated to Mansi again, a program code was developed to replace the Hungarian terms with the already existing translations from the dictionary. Only literals are replaced, definitions and examples are left untouched, so that the linguists can check the actual meaning and can replace them with their Mansi equivalents. The Mansi specialists' role is to check the automatic replacement and to give new term candidates if there is no proper automatic translation.

In this workphase, as there are no synonym dictionaries or thesauri available for the Mansi language, the above-mentioned bilingual student dictionaries are used as primary resources. These dictionaries were designed to be used during school classes, they rarely contain any synonyms, antonyms or hypernyms, and hardly any phrases or standing locutions. (Most of these dictionaries were written by the same authors, thus – besides the inconsistent marking of vowel length – fortunately we do not have to pay special attention to possible contradictions or incoherence.) Hence originates the unbalanced situation in which we are either missing the Mansi translation, either the Mansi definition belonging to the same code, and we are able to present the translation, the definition and the examples of usage only in a few extraordinary instances. The sentences illustrating usage in the synset come from our Mansi corpus, built from articles from the Mansi newspaper called *Luima Seripos* published online semimonthly at `http://www.khanty-yasang.ru/luima-seripos`. In its final version, our corpus will contain above 1,000,000 tokens, roughly 400,000 coming from the online publications and the rest from the archived PDF files.

Even if based on the Hungarian WordNet, the elements of the Mansi WordNet can be matched to the English ones and those of other wordnets since the Hungarian WN itself is paired with the Princeton WordNet (Miller et al., 1990).

### 3.3 Bear language

Another very special problem occurred during wordnet building in Mansi, that is the question regarding the situation of the so called "bear language". The bear is a prominently sacred animal venerated by Mansi, bearing great mythical and ritual significance, and also surrounded by a detailed taboo language. Since the bear is believed to understand the human speech (and also to have sharp ears), it is respectful and cautious 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. The taboo words of this "bear language" may be divided into two major subgroups: Mansi words which have a different, special meaning when used in connection with the bear (e.g. сосыг 'currant' but also meaning 'eye', when speaking of the bear's eyes), and those which may be used solely in connection with the bear (e.g. хащлы 'to be angry', as opposed to кантлы 'to be angry' speaking of a human). Even the word for bear belongs to taboo words and has only periphrastic synonyms like Вōртōлнōйка 'an old man from the forest' etc.

As a first approach, taboo words were included as literals in the synsets because their usage is restricted in the sense that they can solely be used in connection with bears. Hence, first we marked the special status of these literals, for which purpose we applied the note "bear". However, it would have also been practical to well differentiate the synsets that are connected to "bears". This can be realized in many ways: for example, the "bear"-variants of the notions should be the hyponyms of their respective notions, like хащлы 'to be angry', which can be considered as a hyponym of кантлы 'to be angry' speaking of a human. However, this solution is not a perfect one since (i) this is not a widespread method either in WordNets of other languages and therefore it would not facilitate WordNet-based dictionaries and (ii) it is not a true hyponym, that is, a real subtype of their respective notion connected to humans. Finally, we decided to put these notions in separate synsets, which has the advantage that these notions are grouped together and it is easier to do a targeted search on these expressions.

## 4 Results

The manual correction of the automatically translated Basic Concept Set 1 is in progress. Currently, the online xml file contains 300 synsets. These synsets had altogether 410 literals, thus a synset had 1.37 literals in average: this proportion was 1.88 in the original Hungarian WordNet xml file. Concerning the proportion of the two part-of-speech categories, nouns prevail over verbs with 210 nouns (70%), 90 verbs.

Presumably 40% of all lexicon entries are multi-word expressions, regardless of word class or derivational processes. In many case when the Russian word refers to special posts or professional person, the proper Mansi word is a roundabout phrase. For example the учитель 'schoolteacher *masc.*' could be translated as ня̄врамыт ханисьтан хум built up of the element *children-teaching man* , and the feminine counterpart учительница 'schoolteacher *fem.*' as ня̄врамыт ханисьтан нэ̄ from *children-teaching woman*. Though the multi-word expressions are highly variable in their elements, replacing the dedicated parts with synonyms, or adding new ones to enrich the layers of senses. The number of multi-word expressions in this version of the Mansi WordNet is 74, that is 18% of all literals.

Section 3.2 enumerated some challenges about transforming an already existing WordNet to Mansi. Some synsets in the Basic Concept Set also have proved to be difficult to handle. For example, the Mansi language is only occasionally (if ever) used in scientific discourse. Therefore, the terms ‘unconscious process’, ‘physiology’ or ‘geographical creature’ cannot have any Mansi equivalents and therefore can be included in the Mansi WordNet only as non-lexicalized items. The number of such literals is 34, that is 16% of all literals.

## 5 Discussion

Building a wordnet for a minority or endangered language can have several challenges. Some of these are also relevant for dead languages, however, wordnets for e.g. Latin (Minozzi, 2009), Ancient Greek (Bizzoni et al., 2014) and Sanskrit (Kulkarni et al., 2010) prove that these facts do not necessarily mean an obstacle for wordnet construction. Here we summarize the most important challenges and how we solved them while constructing the Mansi wordnet.

### 5.1 Wordnet construction for minority and endangered languages

First, linguistic resources, e.g. mono- and multilingual dictionaries may be at our disposal only to a limited extent and second, there might be some areas of daily life where only the majority language is used, hence the minority language has only a limited vocabulary in that respect. As for the first challenge, we could rely on the Mansi-Russian-English-Hungarian dictionary under construction, which is itself based on Mansi-Russian dictionaries (see above) and we made use of its entries in the semi-automatic building process. However, if there are no such resources available, wordnets for minority languages should be constructed fully manually. For dead languages which are well-documented and have a lot of linguistic descriptions and dictionaries (like Latin and Ancient Greek), this is a less serious problem.

As for the second challenge, we applied two strategies: we introduced non-lexicalized synsets for those concepts that do not exist in the Mansi language or we included an appropriate loanword from Russian.

Besides being a minority language, Mansi is also an endangered language. Almost none of its native speakers have been trained in linguistics, which fact rules out the possibility of having native speakers as annotators. Thus, linguist experts specialized in the Mansi language have been employed as wordnet builders and in case of need, they can contact native speakers for further assistance. This problem is also relevant for dead languages, where there are no native speakers at all, however, we believe that linguists with advanced knowledge of the given language can also fully contribute to wordnet building.

### 5.2 Specialties of wordnet construction for Mansi

Wordnet building for Mansi also led to some theoretical innovations. As there is a subvocabulary of the Mansi language related to bears (see above), we intended to reflect this distinction in the wordnet too. For that reason, we introduced the novel relation “bear”, which connect synsets that are only used in connection with bears and synsets that include their “normal” equivalents. All this means that adding new languages to the spectrum may also have theoretical implications which contribute to the linguistic richness of wordnets.

## 6 Conclusions

In this paper, we reported the construction of a wordnet for Mansi, an endangered minority language spoken in Russia. As we intend to make the Mansi wordnet freely available for everyone, we hope that this newly created language resource will contribute to the revitalization of the Mansi language.

In the future, we would like to extend the Mansi wordnet with new synsets. Moreover, we intend to create applications that make use of this language resource, for instance, online dictionaries and linguistic games for learners of Mansi.

## 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

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti, and Wim Peters. 1998. The Linguistic Design of the EuroWordNet Database. *Computers and the Humanities. Special Issue on EuroWordNet*, 32(2-3):91–115.

A.N. Balandin and M.I. Vahruševa. 1958. *Mansijski-russkij slovar' s leksičeskimi paralelljami iz južno-mansijskogo (kondinskogo) dialekta*. Prosvešenije, Leningrad.

Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors. 2010. *Principles, Construction and Application of Multilingual Wordnets. Proceedings of GWC 2010*. Narosa Publishing House, Mumbai, India.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of ancient greek wordnet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1054.

Christiane Fellbaum and Piek Vossen, editors. 2012. *Proceedings of GWC 2012*. Matsue, Japan.

M. Kulkarni, C. Dangarikar, I. Kulkarni, A. Nanda, and P. Bhattacharya. 2010. Introducing Sanskrit WordNet. In *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, Mumbai, India. Narosa Publishing House.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 311–320, Szeged. University of Szeged.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Stefano Minozzi. 2009. The Latin WordNet project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*, pages 707–716.

Heili Orav, Christiane Fellbaum, and Piek Vossen, editors. 2014. *Proceedings of GWC 2014*. Tartu, Estonia.

E.I. Rombandeeva and E.A. Kuzakova. 1982. *Slovar' mansijsko-russkij i russko-mansijskij*. Prosvešenije, Leningrad.

E.I. Rombandeeva. 2005. *Russko-mansijskij slovar'*. Mirall, Sankt-Peterburg.

Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors. 2008. *Proceedings of GWC 2008*. University of Szeged, Department of Informatics, Szeged, Hungary.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, 7(1-2):9–43.

Veronika Vincze, Ágoston Nagy, Csilla Horváth, Norbert Szilágyi, István Kozmács, Edit Bogár, and Anna Fenyvesi. 2015. FinUgRevita: Developing Language Technology Tools for Udmurt and Mansi. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages*, Tromsø, Norway, January.