My Husband, the Dragonfly: The (Non-)Compositionality of Mansi Animal Names

Veronika Vincze[1,2]
[1]University of Szeged
[2]MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

Csilla Horváth
University of Szeged
naj.agi@gmail.com

Abstract

Here we focus on a specific type of multiword expressions, namely animal names in Mansi, an endangered language spoken in Russia. Mostly, animal names inherently include a meaning component related to a family term or a gender-marking word, thus they may show a high degree of non-compositionality. We present our data collected from dictionaries and folk tales of the Mansi language and we also demonstrate that machine learning tools can effectively learn the distinction between lexically male and female animals.

1 Background

In natural languages there are many ways to express complex human thoughts and ideas. This can be achieved by exploiting compositionality, i.e. concatenating simplex elements of language and thus yielding a more complex meaning that can be computed from the meaning of the original parts and the way they are combined. However, non-compositional phrases can also be found in languages, which are complex phrases that can be decomposed into single meaningful units, but the meaning of the whole phrase cannot (or can only partially) be computed from the meaning of its parts. Such phrases are often called multiword expressions (MWEs) and they display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Calzolari et al., 2002), which might pose problems for linguistic processing, especially in language learning and natural language processing (NLP).

Here we focus on a specific type of multiword expressions, namely animal names in Mansi, an endangered language spoken in Western-Siberia. In several cases, animal names inherently include a meaning component related to a family term or a gender-marking word, thus they may show a high degree of non-compositionality. First we describe Mansi people and their cultural background, then we present our data on Mansi animal names. Later, we show that machine learning techniques can be applied to distinguish among inherently male and female animals. In this way, we would like to emphasize that low-resourced languages can also be of interest for lexical semantic research,with special regard to multiword expressions and non-compositionality.

2 Mansi Culture

Although the prestige of Mansi language and culture is rising, the number of Mansi speakers is critically low. Mansi plays limited role in its Russian-dominated, multi-ethnic and multilingual environment, it is heavily affected by the loss of the traditional way of life and rapid urbanisation as well. While the Mansi have been (and in some respect still are) regarded as followers of traditional, nomadic lifestyles, and are expected to live in rural conditions, the majority of the Mansi live in multiethnic urban environment.

The Ob-Ugrians traditionally belonged to two exogamous phratries called *moś* and *por*. The ancestor of the *moś* phratry was considered to be the godmother *kaltaś-ēkwa*, and his son, *mir-susne-χum*, one of the most important gods in the Mansi pantheon. Genealogical divisions of the tribes and families in the *moś* phratry originated, according to legends, from common ancestors, including zoomorphic ones, such as hawk, frog, etc. (Gemuev, 2008). The *por* phratry considered the bear to be its ancestor. Traces of horse worship denote the *moś* phratry's Ugric origin, while bear-cult confirms the *por* phratry's relationship to the assimilated Siberian population.

3 Dataset

Researchers of the Mansi language already compiled some dictionaries of the language about one hundred years ago, which were only lately published (Munkácsi and Kálmán, 1986; Kannisto, 2013). These dictionaries contain words from all the dialects, also from those that are now extinct. There are also some modern dictionaries of the Northern Mansi dialect available (Rombandeeva, 2005; Rombandeeva and Kuzakova, 1982). We relied on these dictionaries when manually collecting animal names form the language data and we also consulted with native speakers. All the data analysed come from the Northern dialect of Mansi.

Partly as the result of the belief in anthropomorphic guardian spirits, the Mansi adorned the most venerated animals with human attributes (e.g. they were believed to understand human talk) and spoke about them with great precaution. The word *ōjka* 'old man, man, husband' or *ēkwa* 'old woman, woman, wife' was often added to the name of ancestors, guardian spirits, their anthropomorphic images were kept in the sacred trunks in the house or in storehouses at the sacred places. In our data, *ōjka* and *ēkwa* can be found attached also to the name of animals with little or no mythological or sacred importance.

All in all, we managed to collect approximately 90 animal names. 17 of them denote bear, the most venerated animal of the Mansi, which reflects that bear is a taboo word and several periphrastic phrases are applied to refer to that animal. These terms often contain names of family members such as 'brother' or 'aunt'.

Animal names typically contain lexical items that refer to gender. In several cases, these names are not compositional, for instance, *tūlmaχ-ōjka* 'wolverine' literally means "thief man". In our dataset, we provide literal translations of these terms, together with their English translation. The dataset is publicly available at `http://rgai.inf.u-szeged.hu/mwe`. A sample from the dataset can be seen in Table 1.

Mansi name	**Translation**	**Literal meaning**
sāli-purn-uj-ōjka	wolf	reindeer-biting animal-man
vōrt ōlnė ōjka	bear	man living in the forest
såwńaχ-nē	jay	jay-woman
χulaχ-ōjka	raven	raven-man
tūlmaχ-ōjka	wolverine	thief-man
urin-ēkwa	crow	waiting woman

Table 1: Mansi animal names.

4 Machine Learning Experiments

In order to test what features are the most deterministic in choosing the lexically selected gender of the animal, we carried out machine learning experiments. We took 29 animal names that were non-compositional (for instance *kēr χum* lit. iron man 'dragonfly') and defined features for them. Our feature set contained mostly features derived from a biological taxonomy (e.g. whether it is a vertebrate, a mammal, a bird etc.), together with features referring to the animal's abilities (e.g. whether it can swim or fly) and other characteristics (whether it is a domesticated animal, or a sacred animal of the Mansi people).

For this dataset, we applied a multilayer perceptron (Bishop, 1995) – as implemented in Weka (Hall et al., 2009) – in a leave-one-out manner, due to the small size of the data. An accuracy of 82.76% could be achieved, which means that 24 animals out of 29 were correctly classified as having female/male meaning components. As for the incorrectly classified animals, lexically male birds were classified as female since most of the birds in the data were lexically female. Conversely, lexically female mammals were often classified as males due to the fact that mammals in the data were mostly lexically male.

Based on the results, we can state that machine learning tools (especially neural networks) can grab and faithfully model the logic behind the apparently random choice of male or female meaning components for Mansi animals.

5 Conclusions

In this paper, we presented a Mansi dataset of non-compositional animal names. They inherently include a meaning component related to a family term or a gender-marking word, and we proved that with machine learning tools it is possible to grab the difference among inherently male and female animal names. We hope that our dataset will raise the interest for the endangered Mansi language and will be useful for Mansi people as well as for NLP researchers interested in endangered, moribund and extinct languages.

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

C. M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, pages 1934–1940.

I. N. Gemuev. 2008. *Mansi mythology*. Akadémiai Kiadó, Budapest.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18.

Artturi Kannisto. 2013. *Wogulisches Wörterbuch*. Kotimaisten Kielten Keskuksen Julkaisuja, Helsinki.

Bernát Munkácsi and Béla Kálmán. 1986. *Wogulisches Wörterbuch*. Akadémiai Kiadó, Budapest.

Evdokija Ivanova Rombandeeva. 2005. *Russko-mansijskij slovar'*. Mirall, Sankt-Peterburg.

Evdokija Ivanova Rombandeeva and Evdokija Aleksandrova Kuzakova. 1982. *Slovar' mansijsko-russkij i russko-mansijskij*. Prosvešenije, Leningrad.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*. Mexico City, Mexico, pages 1–15.