On the Road to a Dialect Dictionary of Khanty Postpositions

Zsófia Schön
Ludwig Maximilian University of Munich
Institute for Finno-Ugric Studies
zsofia.schoen@gmail.com

December 17, 2014

Abstract

This paper aims to present the first steps of a corpus based dialect dictionary of postpositions in several Khanty dialects and subdialects. Based primarily on specifically elicitated data from more than fifty informants, this ongoing project focuses not only on the semantic properties of this part of speech in Khanty, but also on the morphology and combinatorics as exhibited by (sub)dialectal microvariation. Special attention is paid to two of the Northern dialects – Kazym and Shuryshkary Khanty – and to one of the Eastern dialects – Surgut Khanty.

The lexicon entries have been compiled according to TEI P5 guidelines in XML format, while the corpus data is stored in a MySQL database. A web application combining the lexicon with the corpus data, sound files, annotations and metadata is currently under construction.

As a multilingual dialect dictionary of Khanty postpositions, this project hopes to fill a gap in current research on Khanty: namely the lack of easily accessible digital dictionaries. It is designed to be a pilot project for forthcoming digital Khanty dictionaries.

1 Introduction

Khanty (Ostyak) language is spoken in North-West Siberia along the river Ob and its tributaries. According to the 2010 Russian census, some 30,000 Khanty are living there, but only about 20 % of them are native speakers.

Belonging to the Ob-Ugric branch of the Finno-Ugric language family, Khanty demonstrates the highest dialectal variability within this language family. Most of these dialects are highly endangered.

I hold the opinion that despite the relative spatial proximity of the dialects and subdialects of Khanty, even the lexicon of a grammatic category, like the one of postpositions, shows a level of variation so high, that the Khanty language cannot be considered a homogeneous entity.

Postpositions are considered a part of speech which marks grammatical function [1]: 224. They can be categorised either as a homogeneous closed word class or as a heterogeneous open word class – depending on the grade of grammaticalisation, e.g. nouns with case suffixes used as postpositions. In Khanty dialects, postpositions occur only with an antecedent at the left which can be a substantive (always in nominative), a pronoun, an adverb, an adjective, a numeral or a participle. The antecedent and the postposition are inseparable and form an adverbial construction. The postpositions themselves can be variable or invariable, i.e. if they take possessive suffixes or not. (For a detailed description and categorisation of Khanty postpositions, see [2].)

1.1 About the Project

This ongoing lexicological project originates from my doctoral thesis about postpositional constructions in several Khanty dialects [2] which was started in October 2009 and completed in July 2014. It contains, as an appendix, a *Dialect Dictionary of Khanty Postpositions*, formatted from XML into LaTeX using XSLT, as well as selected sample data from the partly tagged and analysed corpus.

The dictionary entries have been compiled in XML according to the TEI P5 guidelines, the corpus data is stored in a MySQL database. The corpus consists of approximately 7,900 sentences containing postpositions, where the postpositional constructions have been glossed and tagged.

The goal of the current project is to present the material and lexicon in a digital and freely accessible format on the web. The project is supported by the IT Group for the Humanities at the Ludwig Maximilian University of Munich, particularly by Christian Riepl and Stephan Lücke as database designers, Gerhard Schön as developer of the web application, and myself working as lexicographer.

1.2 About the Khanty Dialect System

Although there is a large number of studies on Khanty dialectology (see e.g. [3], [4], [5], [6]), no consensus has been established so far about terminology or structure of the Khanty dialect system. Different regroupings and dialect names can be found in the literature, where either phonetic or morphologic aspects have been taken to classify the Khanty dialects (cf. e.g. [7], [8], [9], [10], [11]).

In this paper, the term *dialect* is used according to the definiton of Chambers and Trudgill: a dialect "refers to varieties which are grammatically (and perhaps lexically), as well as phonologically different from other varieties" [12]:5.
Subdialects, however, form smaller entities, are even more concrete than dialects and always mutually intelligible. The minor differences between subdialects appear only in the lexicon or on a phonetic level.

Existing heterogenous approaches, diachronic tendencies and endangerment of dialects necessitate the establishment of a coherent classification of Khanty dialects not only on a dialectal, but also on a more concrete, subdialectal level.

Therefore, I propose a dialect system with a set of subdialects; see Table 1, described in my doctoral thesis ([2]: 9–45), although the number of subdialects could increase as a result of further research. The compilations were made on the basis of all accessible Khanty dialect dictionaries, especially [13] and [14], of secondary literature, as well as of own field notes and field observations.

2 The Digital Dialect Dictionary of Khanty Postpositions

In print, there exist excellent dictionaries of Khanty dialects, like the DWS [13] or the SVH [14]. In digital form, however, easily usable and accessible dictionaries are still missing.

The current project aims at changing this by creating an easily searchable multilingual online dictionary with a user-friendly web interface.

The lexicon entries were compiled on the one hand from a synchronic point of view, which is guaranteed by the corpus data collected between 2010 an 2012, and on the other hand from a diachronic point of view, which is ensured by information provided by the two aforementioned lexical resources.
Basically, the lexicon entries are sorted alphabetically, where the vowels are judged one by one, unlike the Finnougric tradition which would treat all vowels as one entity.

The digital lexicon entries will be enriched by sound files, annotations, as well as metadata about informants, dialects and subdialects, circumstances of the collection etc.

	Eastern dialect group	Southern dialect group	Northern dialect group
Dialect *Subdialect*	***Vakh-Vasyugan dialect***	***Irtysh dialect***[1]	***Middle-Ob dialect*** Keushi subdialect Muligort subdialect Nizyam subdialect Sherkaly subdialect
Dialect *Subdialect*	***Surgut dialect*** Pim subdialect Tremyugan subdialect Tromagan subdialect Agan subdialect Yugan subdialect Likrisovskoe subdialect		***Kazym dialect*** Lower Kazym subdialect Middle Kazym subdialect Upper Kazym subdialect
Dialect *Subdialect*	***Salym dialect***		***Beryozov dialect*** Tegi subdialect
Dialect *Subdialect*			***Shuryshkar dialect*** Poslov subdialect Shuryshkar subdialect Muzhi subdialect Synya subdialect
Dialect *Subdialect*			***Obdorsk dialect*** Poluy subdialect Sob subdialect

Table 1: Khanty dialect and subdialect system compiled from [2]: 45.

[1] Since the extent of divergation of the around twenty-six subdialects of the Irtysh dialect is unclear, and as it is impossible to say how many of them can be really called subdialects, because they are all extinct, I refrain from naming them here.

The headwords of the lexicon entries are organised with the help of abstract lemmata, which were elaborated by referencing to the phoneme systems of the dialects from a synchronic point of view. This was necessary because of the strong divergence between the phoneme systems, and makes it possible to organize the very same postposition from different dialects into one entry without favouring one form from one dialect over another, and without relying on the knowledge of the dictionary user about the historical phonology of Khanty.

At present, the languages of the dictionary are seven Khanty subdialects as well as German, but it is intended to add more Khanty (sub)dialects from mostly written sources and a translation into Russian.

2.1 The Corpus

The corpus data stored in a MySQL database was collected by myself in Siberia during four field trips, all together six months, between 2010 and 2012 from over fifty native speakers of three dialects and seven subdialects, in particular of the Tromagan and Yugan subdialects of the Surgut dialect, all three subdialects of the Kazym dialect and all two abovementioned subdialects of the Shuryshkary dialect.

The material consists of a total of over 7,900 sentences containing postpositions, which were elicitated mainly with the help of Russian stimuli. Even if this method has its limitations (e.g. the data is not suitable for analysing word order), it allowed me to assemble a heterogenous collection of postpositions in a short time from a high number of informants with subdialectal microvariation. The free translations show a large range of usage of the different structures in contemporary spoken language.

The postpositional constructions have been glossed and tagged according to the Leipzig glossing rules.

Concerning the transcription of the data, a transcription in IPA was elaborated – for Kazym and Surgut Khanty (sub)dialects, this was already done in the EUROBABEL project *"Ob-Ugric languages: conceptual structures, lexicon, constructions, categories"*, followed by the subdialects of Shuryskary Khanty during my work on the doctoral thesis. The transformation of the data into cyrillic orthographies for the different (sub)dialects will be planned at a later stage.

2.2 Structure of the Lexicon Entries

As mentioned before, the lexicon entries have been compiled according to TEI P5 guidelines in XML format. Special consideration has been given to the morphology of the postposition, its combinatorics and the highlighting of the (sub)dialectal level of information, with dialect abbreviations from the two aforementioned dictionaries and my collection.

The structure of the lexicon entries looks as follows, e.g. in the case of the headword *JUP-*:

The headword is immediately followed by the etymology of the postposition as given in the dictionary DWS [13]:

```
<entry>
    <form type=”lemma”>JUP-</form>
    <etym>Grammatikalisiert aus Substantiv mit der Bedeutung
    ’Hinterseite (eines lebenden Wesens)’.
       <bibl>DWS 328</bibl>
    </etym>
```

Then the forms of the postposition itself are given along with specified usage and details about the subdialect in which the forms of the postpositions occur:

```
<gramGrp>
   <case n=”I”>In Lokativ</case>
</gramGrp>
<form type=”inflected”>jupijən
   <lang>KAZ_KAM[2]</lang>
   <lang>KAZ_KAK</lang>
   <lang>KAZ_KAO</lang>
   <lang>SHU_POS</lang>
   <lang>SHU_SYN</lang>
</form>
<form type=”inflected”>jupe-
   <gram xml:lang=”KAZ-KAM”>
      vor Possessivsuffix (3SG)
   </gram>
   <gram xml:lang=”KAZ-KAO”>
      vor Possessivsuffix (3SG)
   </gram>
</form>
```

[2] In the example cited here, the following (sub)dialects and abbreviations occur: Lower Kazym (KAZ_KAM), Middle-Kazym (KAZ_KAK) and Upper Kazym (KAZ_KAO) subdialects of the Kazym dialect; Poslov (SHU_POS) and Synya (SHU_SYN) subdialects of the Shuryshkar dialect as defined in [2] (SZS). From DWS, some abbreviations have been taken as well: Kazym dialect (Kaz.), Kazym dialect collection of Wolfgang Steinitz (KazSt.), Shuryshkar dialect (Š), Synya dialect (Sy.), collection of József Pápay and Beke Ödön (PB).

```
<form type=”inflected”>jupɛ-
   <gram xml:lang=”KAZ-KAM”>
      vor Possessivsuffix (3PL)
   </gram>
   <gram xml:lang=”KAZ-KAK”>
      vor Possessivsuffix (3SG)
   </gram>
</form>
```

Once the forms of the postposition are enumerated, the meanings can be worked out. Each meaning is provided not only with details about the subdialect in which it occurs, but also with information about the combinatorics of the meaning, its sources in the two aformentioned dictionaries or my collection. If documented in the corpus, the meaning is illustrated by sample sentences and their metadata:

```
<sense n=”1”>
   <gramGrp>Mit Nomen</gramGrp>
   <lang>Š</lang>
   <lang>KazSt.</lang>
   <lang>Sy.</lang>
   <lang>PB</lang>
   <def>nach, hinter</def>
   <ref>DWS 328</ref>
</sense>
<sense n=”1a”>
   <gramGrp>Mit Personalpronomen</gramGrp>
   <lang>Kaz.</lang>
   <lang>Sy.</lang>
   <lang>KAZ_KAM</lang>
   <lang>KAZ_KAO</lang>
   <def>hinter jemandem
      <eg xml:lang=”KAZ-KAO”>ɬuβ jupe=ɨ=ən βɛɬpəsti mantaɬən
      aːmpəɬ ʃɵtɬ</eg>
      <eg xml:lang=”GER”>Hinter ihm, wenn er jagen geht,
      geht sein Hund.</eg>
      <bibl>ETM 202/028</bibl>
   </def>
   <ref>DWS 328</ref>
   <ref>SZS</ref>
</sense>
```

The entry is concluded either with the last meaning or with possibly already grammaticalised phrases:

```
    <dictScrap>
       sʲi ~
       <desc>danach, dann</desc>
       <lang>Š</lang>
       <lang>PB</lang>
       <lang>KAZ_KAM</lang>
       <lang>KAZ_KAO</lang>
       <lang>SHU_POS</lang>
       <lang>SHU_SYN</lang>
       <ref>DWS 328</ref>
       <ref>SZS</ref>
    </dictScrap>
</entry>
```

3 Further steps

The material is now being transferred into a web application combining the XML TEI P5 documents of the lexicon and the MySQL data of the corpus. Each headword will come with the full lexicon entries as described above, with hyperlinks from each (sub)dialect, category of form, usage and meaning to all corresponding occurrences in the corpus, along with sound files, annotations and metadata. Aside from a simple search over the most common attributes, an advanced search mode will allow for a variety of combinations of search terms and clauses. A contact form will be provided for feedback and suggestions.

References

[1] R.M.W. Dixon. *Basic Linguistic Theory. Volume 1. Methodology.* Oxford University Press, 2010.

[2] Zsófia Schön. *Postpositionale Konstruktionen in chantischen Dialekten.* PhD thesis, Ludwig-Maximilians-Universität, München, 2014.

[3] Ф.М. Лельхова. Шурышкарский диалект (сынский говор) в системе диалектов хантыйского языка. In *Актуальные проблемы разработки учебно-методических комплектов по хантыйскому и мансыйскому языкам;*

литературе и культуре: Материалы окружной научно-практической конференции, посвяшенной 70-летнему юбилею Е.А. Нёмысовой. 3–5 мая 2006 года, pages 85–90, Ханты-Мансийск, 2007. Полиграфист.

[4] Irina A. Nikolaeva. On the state of dialectological studies of ostyak in the ussr in connection with the program ”dialectologia uralica”. *Mitteilungen der Societas Uralo-Altaica*, 10:59–61, 1990.

[5] Erhard F. Schiefer. Kriterien zur klassifizierung der dialekte des ostjakischen. *Dialectologia Uralica. Veröffentlichungen der Societas Uralo-Altaica*, 20:257–281, 1985.

[6] Mária Sipos. Északi hanti: nyelvjárási kontinuum vagy nyelvjáráscsoportok? *Folia Uralica Debreceniensia*, 20:251–258, 2013. http://finnugor.arts.unideb.hu/fud/fud20cikkek/14_sipos.pdf.

[7] Daniel Abondolo. Khanty. In Daniel Abondolo, editor, *The Uralic Languages*, pages 358–386. Routledge, London – New York, 1998.

[8] László Honti. Die ostjakische sprache. In Denis Sinor, editor, *The Uralic Languages. Description, History and Foreign Influences. Handbuch der Orientalistik*, pages 172–196. E. J. Brill, Leiden – New York – København – Köln, 1988.

[9] Вольфганг Стейниц. Хантыйский (остяцкий) язык. In *Ostjakologische Arbeiten*, volume IV, pages 5–62. Akadémiai Kiadó – Akademie-Verlag, Budapest – Berlin, 1937/1980.

[10] Wolfgang Steinitz. *Ostjakische Grammatik und Chrestomathie mit Wörterverzeichnis.* Otto Harrassowitz, Leipzig, 1950. 2., verbesserte Auflage.

[11] Н.И. Терёшкин. Хантыйский язык. In В.И. Лыткин and К.Е. Майтинская, editors, *Финно-угорские и самодийские языки. Языки народов СССР*, volume III, pages 319–342. НАУКА, Москва, 1966.

[12] J.K. Chambers and Peter Trudgill. *Dialectology.* Cambridge University Press, second edition, 1998.

[13] Wolfgang Steinitz et al. *Dialektologisches und etymologisches Wörterbuch der ostjakischen Sprache – Диалектологический и этимологический словарь хантыйского языка.* Akademie-Verlag, Berlin, 1966–1993.

[14] Н.И. Терёшкин. *Словарь восточно-хантыйских диалектов.* НАУКА, Москва, 1981.