LINGUISTIC VARIATION AND MINOR LANGUAGES CORPORA: A CASE STUDY OF MANSI DIALECTS[1]

Zhornik D. O. (daria.zhornik@yandex.ru), Moscow State Lomonosov University, Moscow, Russia

Sizov F. O. (f.sizov@yandex.ru), Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

Annotation: This paper presents a multidialectal corpus of the Mansi language and introduces our methods of solving the problem of dialectal variation. Mansi (< Ob-Ugric < Finno-Ugric < Uralic) is a relatively poorly documented language. So far, there is not a single tolerable resource which would contain annotated texts in various Mansi dialects. Our corpus is an attempt at creating such a resource. The data to be found in the corpus are diverse both regarding the dialect they belong to and the time when they were recorded. Texts in extinct Mansi dialects (Southern, Eastern, Western) may date as far back as the 1840s, while Upper Lozva texts and audio recordings were gathered in 2017-2018. Because of this diversity, the problem of linguistic variation and its support in the corpus is crucial for us. The data which exhibit both linguistic variation and heterogeneity of writing systems need to be processed uniformly. The complex task of variation processing is solved separately at each step of corpus building: optical character recognition (OCR) during preliminary processing of printed materials, morphological annotation and search implementation based on the Tsakorpus platform.

Keywords: corpus linguistics, language documentation, minor languages, Finno-Ugric languages, morphological analyzer, computational linguistics

ЯЗЫКОВАЯ ВАРИАТИВНОСТЬ И КОРПУСА МАЛЫХ ЯЗЫКОВ: ИССЛЕДОВАНИЕ МАНСИЙСКИХ ДИАЛЕКТОВ

Жорник Д. О. (daria.zhornik@yandex.ru), Московский государственный университет имени М. В. Ломоносова, Москва, Россия

Сизов Ф. О. (f.sizov@yandex.ru), Институт языкознания РАН, Москва, Россия

Аннотация: Статья представляет многодиалектный корпус мансийского языка, а также способы решения проблемы диалектной вариативности при его разработке. Мансийский язык (обско-угорские < финно-угорские < уральские) относительно плохо документирован, и на данный момент не существует ни одного удовлетворительного ресурса, объединяющего в себе аннотированные тексты на многочисленных мансийских диалектах. Представленный корпус призван заполнить эту лакуну. Данные, содержащиеся в корпусе, многообразны как с точки зрения диалектных вариантов, так и по своей временной принадлежности. Тексты на вымерших мансийских диалектах (южные, восточные, западные) датируются 19-20 веками, в то время как тексты на верхнелозьвинском диалекте были записаны в 2017-2018 годах. В связи с этим разнообразием особенно важной представляется проблема вариативности и её поддержки в корпусе. Необходимо единым образом обрабатывать данные, проявляющие как языковую вариативность, так и неоднородность систем записи. Сложная задача вариативности решается отдельно для каждого этапа построения корпуса: оптического распознавания текстов при предварительной обработке печатных материалов, морфологической разметки и реализации поиска в корпусе на основе платформы Tsakorpus.

Ключевые слова: корпусная лингвистика, языковая документация, малые языки, финно-угорские языки, морфологический анализатор, компьютерная лингвистика

[11] The current research was funded by the RFBR project 18-012-00833 A "Dynamics of phonetical and grammatical systems of Ob-Ugric languages".

1. Introduction

Small-sized corpora of weakly documented (and predominantly non-written) languages constitute a specific domain within the vast field of corpus linguistics (see [Ostler 2008]), much in between language documentation in the traditional sense and modern corpus-oriented research. Gathering linguistic data and integrating this material into a readily usable corpus are equally non-trivial tasks with many pitfalls and challenges. One of the main issues here is processing linguistic variation which may be notoriously prolific in case a written norm (and often any written practice at all) is absent.

Unfortunately, we face a very limited number of generally accepted authoritative computer tools for this type of linguistic data (i.e., for weakly documented languages with a considerable degree of internal variation), though some efforts in this direction are worth noticing (cf., for example, [Simon & Mus 2017]). It seems that the present stage of corpus linguistic research should privilege specific case studies which may be further compared to provide the most efficient solutions. The present paper will discuss one such case related to computer tools for documenting the Mansi language (Ob-Ugric < Finno-Ugric < Uralic).

Mansi (Vogul), with its closest relative Khanty (Ostyak), are two endangered representatives of the Ugric branch (which includes the more remote Hungarian as well); their speakers dispersedly inhabit a large area in the northern part of Western Siberia. Strictly speaking, Mansi and Khanty are not just two different languages, but a complex continuum of more or less closely related local varieties; this is especially true for Khanty which exhibits significant geographical variation as against East and West dialect groups. The Mansi area is much smaller and can be visually represented as a narrow discontinuous strip on the western and southwestern border of Khanty zone.

Overall, Mansi is more heavily endangered than Khanty and less documented at the same time. Three of four main Mansi dialect groups can be regarded as extinct nowadays, and only one (namely, the Northern) survives, mainly among the speakers of the elder generations. It makes the documentation of still existing Mansi varieties an urgent task. Remarkably, the authors of the present paper recently came across a unique “linguistic island” of an entirely preserved variety of Northern Mansi, located in three separate hamlets on the Upper Lozva river. This subdialect within the Northern dialectal zone (with approx. 100 speakers) seems to be barely reported in existing documentation on Mansi (see [Zhornik, Pokrovskaya 2017]).

A few Upper Lozva texts are found among A. Reguly’s materials in A. Kannisto’s “Wogulische Volksdichtung”, which date back as far as the 1840s and the 1900s respectively. There is also a small collection (consisting mainly of wordlist elicitation) of modern Upper Lozva recordings gathered by a Hungarian researcher Gábor Székely within his work for the ELAR SOAS project (https://elar.soas.ac.uk/Collection/MPI931196). Apart from that, almost no data on the Upper Lozva variety can be found. The paper [Székely 2012] summarizes all available information on the previous fieldwork among the Upper Lozva Mansi. In 2016, Tatiana Bakhtiyarova, a native speaker of the Upper Lozva dialect and a current member of our project, has published (together with Svetlana Dinislamova) a dictionary of her home language variety. [Bakhtiyarova, Dinislamova 2016] is an extremely valuable resource for our project; thus far, we have converted the information it contains into a digital database.

Due to the lack of oral data, the recorded, transcribed and annotated texts from Upper Lozva play a crucial role in our multidialectal corpus. However, our project encompasses not only the Upper Lozva dialect but all Mansi varieties, for which at least some documentation exists. As Mansi exhibits a considerable level of dialectal variation, the main problem we have to tackle is, as mentioned above, the processing of such variation.

In the remainder of this paper, we are going to describe the main strategies of supporting cross-dialectal variation that apply on different stages of corpus building:

- processing using OCR tools

- morphological annotation
- implementing search interface.

Accordingly, the paper consists of 6 sections. **Section 2** briefly discusses available computational resources for Uralic languages in general and Mansi in particular. **Section 3** presents our project on documentation of the Mansi language, while **section 4** elaborates on the stages of preliminary data processing for our corpus. **Section 5** introduces the morphological parser AmpEngine specially developed for our project. The concluding section offers a summary of all the topics mentioned above.

2. Existing sources

As for Uralic studies in general, there is a growing understanding concerning the need for computational tools for compiling text databases and full-fledged corpora and developing means for automatic morphological and syntactic analysis. Some significant publications on this subject are [Moshagen et al. 2014], [Gerstenberger et al. 2017], [Simon, Mus 2017], to mention but a few.

Compared to other Uralic languages, digitized Mansi texts (which are our primary concern) appear rather scarcely. Some texts can be found on websites of the newspaper "Lūima Sēripos" (http://www.khanty-yasang.ru/luima-seripos) and of the Ob-Ugric institute of applied research and development (https://ouipiir.ru). Both resources are based in Khanty-Mansijsk, Russia and prove to be useful for our purposes. However, these websites only provide plain texts, that is, no glossing or annotation for them is available.

On the other hand, a corpus of 272 (to some extent) annotated Mansi texts exists as part of the Ob-Ugric Database (OUDB) founded by Elena Skribnik and based in Munich. For its description, see http://www.babel.gwi.uni-muenchen.de, as well as [Schön, Wisiorek 2016]. Another resource is "Languages under the influence" project, presented in [Simon, Mus 2017]. However, the corpus described in the latter contribution is not available and hence not open for evaluation.

All these resources appear to be "work-in-progress" and consequently, are characterized by some degree of incompleteness. First, the number of texts is limited: either we deal with the old records made at the turn of the 20th century by Finnish and Hungarian explorers or with a small subcorpus of written literary texts from the Soviet and post-Soviet period (mainly, periodicals — even fiction is underrepresented). Audio and multimedia corpora are practically non-existing, and, more importantly, there is virtually no documentation of Mansi dialects. Second, even if the available texts are glossed, these glosses only include inflectional morphological categories. Neither derivational affixes nor periphrastic forms (rather numerous in the Mansi verbal system) are analyzed. Finally, the existing tools are poorly adapted for tackling dialectal variation of various types. It is this latter objective which is central to our project aimed at the extensive corpus documentation of the Mansi linguistic continuum.

3. Project on Mansi Documentation

In 2017, a new project for in-depth studying of the Mansi language was launched (see http://digital-mansi.com/ for more information). Two primary goals of the project are documentation and description of the Mansi language. Consequently, we focus on creating a multidialectal text corpus on the one hand, and performing fieldwork among the speakers of the Upper Lozva dialect, on the other hand.

The corpus presented on the website is going to consist of a significant number of subcorpora (currently under construction) distributed according to dialectal areas. Several subcorpora will also feature standard written Mansi: newspapers, fiction, Bible translations, and so forth. Each subcorpus (or collection thereof) exhibits its own lexicon, morpheme list and grammar rules.

The written texts for the corpus originate from different sources, such as books and newspaper articles in standard Mansi, as well as fieldwork data from various Mansi dialects recorded in the 19th and 20th centuries by Arturi Kannisto, Antal Reguly, Béla Munkácsi, Nikolay Chernetsov and other researchers.

Dialectal subcorpora contain texts gathered from speakers of Southern, Eastern, Western and Northern Mansi varieties.

Our corpus contains only one multimedia subcorpus (provided with audio files), as most of the Mansi dialects are extinct and have no audio recordings whatsoever. The project's participants recorded the audio materials used in the corpus during their fieldwork in the Upper Lozva region in 2017 and 2018. So far, we have recorded 4 hours of Upper Lozva Mansi texts, produced by 13 speakers (both male and female, aged approx. 20 - 70). These recordings are being transcribed and segmented into clauses in ELAN. Thus, we acquire audio markup for the multimedia subcorpus, which is afterwards synchronized with morphological annotation.

Since the data included in the corpus are highly diverse, the main problem we are forced to tackle during corpus development is cross-dialectal variation support. Most of the project's resources are designed for solving these issues. Unfortunately, available morphological parsers are unable to cope with this task on a full scale. Given this, we have developed a specialized morphological analyzer AmpEngine based on the models borrowed from the universal morphological analyzer UniParser (see [Arkhangelskiy et al. 2012] for more detail). The main goal of AmpEngine is to capture cross-dialectal variation, as well as to integrate written and oral texts within the same automated system.

4. Preliminary processing of Mansi texts

The first step when preparing printed sources for including in the corpus was to apply Optical Character Recognition (OCR) techniques to the images. For the present purpose, the OCR software should support cross-dialectal variation, so it should be able to recognize numerous character sets, which may represent various writing systems depending on the author, the time period, the described dialect, and so forth. Supporting various fonts is also required for this end.

The scanned images were recognized and converted to the plain-text format with the Mansi module for open-source Tesseract OCR Engine, which was specially developed and trained by the authors. It may be used to recognize paper-based literature written in Mansi at different times. Currently, optical recognition was performed only for texts in standard Mansi.

However, we intend to add subcorpora containing texts in non-written Mansi dialects as well. These texts were recorded by different researchers, who designed various transcription methods, so they exhibit a high level of variation in terms of writing systems. Researchers, who recorded Mansi texts in the 19th and early 20th centuries, aimed at a high degree of phonetic accuracy and thus the texts transcribed by them contain numerous symbols and diacritics, which are not found in Unicode and are often used inconsistently. During the process of optical recognition of such texts, we are forced to decide which Unicode symbols correspond to the ones found in texts most closely. Due to this problem, we must put considerable effort into learning OCR module performed entirely manually, as far less content of model data may help Tesseract to accelerate the process. The accuracy of optical recognition of standard Mansi texts equals 84%. OCR processing for dialectal texts is still in progress, so the accuracy thereof cannot be estimated yet. After optical recognition, a group of systematic mistakes produced by the engine is improved via regular replacements with the help of *sed* stream editor[2]. Residuary mistakes may be found and corrected manually.

After optical recognition, intermediate correction of recognized texts is performed. It includes automatic searching and correction of rare mistakes identified in the texts with the help of spell-checking algorithms. If there is more than one language in the text, we take into account the language of the token (generally, Russian vs. Mansi) to enhance correction efficiency and to escape improper corrections. In other words, if we can determine the language of the token, we are more likely to avoid confusion of spell-checking methods for different languages.

[2] https://www.gnu.org/software/sed/manual/sed.html

Thus, drawing on scanned versions of all the available Mansi dictionaries, that is [Chernetsov, Chernetsova 1936], [Balandin, Vakhrusheva 1958], [Rombandeeva, Kuzakova 1982], [Munkacsi, Kalman 1986], [Rombandeeva 2005], [Kannisto 2014], [Bakhtiyarova, Dinislamova 2016], we have created several databases. As a result of OCR processing of the above dictionaries, we acquired plain text versions, which, in turn, were parsed using context-free grammars combined with regular expressions (via Perl6 grammars and ANTLR4). The number of entries currently used in the parser is 10262.

5. Morphological parser and variation support

We have chosen Tsakorpus (developed by Timofey Arkhangelskiy, see https://bitbucket.org/tsakorpus/tsakonian_corpus_platform) as our corpus platform, as it answers our purposes exceptionally well. Notably, Tsakorpus provides integration of audio and video files, as well as advanced variation support. The automatic phonological and morphological parser for Mansi dialectal texts is based on the universal AmpEngine tool, which has been developed by Fedor Sizov within the project on the Mansi language documentation.

The testing of the morphological parser for the Mansi language has been executed based on the newspaper "Lūima Sēripos", which is openly accessible on the website http://www.khanty-yasang.ru/luima-seripos. The results of parsing have been compared with the previously performed manual annotation. The testing involved 7 newspaper articles, containing in total 2317 wordforms. The latest version of the analyzer reached 87% accuracy. Typically, the words, for which the parser was unable to provide correct analysis, are all either toponymes, Russian calques or neologisms. These types of words are almost never found in dictionaries.

Below we present examples of sentences in standard Mansi as processed by the analyzer AmpEngine (all examples are taken from the "Lūima Sēripos" newspaper). For the sake of clarity, we have manually added English translations of each sentence.

(1) *Тāн* *пуссын* *аквхурип* *тāрвитыӈ* *вāрмаль* *ōньщ-ēг-ыт.*
they(pl.) all the.same difficult task have-PRS.IND-3PL
sinew
'They all had the same difficult task'.

(2) *Ты* *нэ̄пак* *тāрат-ан* *мāгсыл* *олн* *тāн* *тāнки* *сэ̄л-э̄г-ыт.*
this paper let.out-PTCP.PRS for money they(pl.) they(pl.)-EMPH earn-PRS.IND-3PL
sinew
'This newspaper is published with the money they earn themselves'.

For the first two instances of the word *тāн*, the parser returns two concurrent representations — a pronoun (they(pl.)) and a noun (sinew). In these cases, the rules of homonym gradation indicate the pronoun as the most likely candidate. In the third case, the word *тāн* is followed by the morpheme *ки* and is thus only marked as a pronoun, despite the availability of concurrent representations for *тāн*. This is due to the distribution of the marker *ки*, which can only be attached to the pronoun, as stated by the filter conditions.

AmpEngine supports various inflection and word formation patterns. Moreover, AmpEngine can process compounds occurring in the dictionary data and integrate them with analytic word formation models (a lot of these models are listed in [Rombandeeva 1973]), which is of utmost importance for parsing Mansi texts (very rich in periphrastic constructions). The use of AmpEngine as a morphological analysis platform is conditioned, in the first place, by these advantages, which are all crucial for the goals of our project. Still, one of the main goals is providing support of writing system variation. It should be emphasized that both dialectal and standard Mansi texts use different writing systems, as each editor or author employs their own strategy of language encoding. Accordingly, there

is no one-to-one correspondence between them, which is obviously problematic for text processing tools. Thus, AmpEngine can be used as a base for a morphological parser, which would suit the requirements of our project.

Lack of uniformity among different writing systems has already been noticed by some researchers working with languages exhibiting a high level of dialectal variation, see, for example, [Jarrar et al. 2014]. The authors of this paper suggest using CODA (Conventional Orthography for Dialectal Arabic) as a single format for recording texts in different Arabic dialects. All original texts are to some extent unified to correspond to this writing system, although it allows some degree of dialectal variation. We find another strategy for solving the problem under discussion in [Simon, Mus 2017]. The authors propose that regular matches for replacements in some conversion directions be created. However, we do not find these solutions entirely appropriate for our purposes, as they cannot correctly reflect the individual characteristics of various Mansi dialects. Therefore, we propose to solve the problem by developing a converting scheme, which would enable free switching of various systems.

As our corpus exhibits various dialects and writing systems, variation processing during a single query is required. For example, while searching for a word, the corpus engine should also be able to find its dialectal variants (or different writing options). While searching for the Northern Mansi word sus 'flea', we expect the corpus to return its Western (*šuš*), Eastern (*šonš*) and Southern (*šoš*) variants. Or, if we search for the standard Mansi word *lātəŋ* 'language', we would like to see the results containing different writing options of this word (found in texts published in different time periods and/or by different authors), such as *laatəŋ, латыӈ, лāтыӈ, лаатынг* and, possibly, a few others. However, the recordings from the Upper Lozva dialect should be added to the corpus in the same system that was used during the fieldwork, as this system was specifically designed for this dialect and reflects its characteristic features in a more precise way (for example, the sounds [i] and [ə] are properly represented in this transcription system, while in standard Mansi orthography they are written with the same letters, namely, *и* and *ы*) . In that case, we propose to use a back-converting action applied to the results of the primary analysis.

In some cases, the orthographic system of standard Mansi seeks to achieve uniformity in language representation and omits specific features of dialectal pronunciation, in contrast to other transcription systems. For example, the Mansi word for 'new' is written in standard orthography only as *лильпи*, the first sound being [l'], while in Upper Lozva Mansi the first sound may be [j], so we would expect the variant *йильпи* as well. To account for such simplification, the layer for symbol conversion for the parsed text may be designed to include all known transcription systems. If the original transcription is more detailed than the writing system to which we would like to switch (e.g., from IPA transcription of dialectal speech to standard orthography), several original symbols merge into one during the conversion process.

Currently, our corpus exhibits two writing systems, namely that of the standard Mansi language and the transcription system used for the texts recorded during fieldwork among the Upper Lozva Mansi. Our parser provides free switching between these two systems: symbols pertaining to both are matched to each other and, if more than one symbol in one writing system corresponds to one symbol in another, the parser applies rules of conversion. These rules consider the environment in which the symbol in question appears. For example, the symbol *a* used our fieldwork transcription may correspond to Cyrillic *а* or *я*: *я* is chosen when the (Latin) a follows a palatalized consonant (in our transcription, palatalization is marked with an apostrophe) or the consonant *j*. In all other cases, a is chosen as a corresponding symbol. We plan to extend our conversion system to all writing systems that will be added to the corpus.

AmpEngine allows us to support not only writing system variation but also lexical and grammatical cross-dialectal diversity. To add Mansi to AmpEngine as a supported language, we developed a new

container[3] based on [Rombandeeva 1973] grammar, which describes inflection and word formation models of the standard Mansi language. Later, we plan to build on the [Rombandeeva 1973] container to develop Mansi dialectal containers.

The container using the grammar [Kulonen 2007] and supporting the Eastern Mansi dialect will be created as soon as enough texts in this dialect are collected and made available. To provide better support of this dialect, AmpEngine can be updated via the addition of new morphological data (for example, list of morphemes and rules of their ordering) and restriction of the existing rules, if some of them are not applicable to the dialect in question. During two-in-one container (basic[4] and attached) merging, conflict situations may emerge. In such cases, a particular element (for example, a morpheme) may cause contradiction between two containers. For example, if in standard Mansi the diminutive marker *-riɛ* is usually attached to nouns and in Eastern Mansi it exhibits transcategorial behaviour by combining with verbal stems, it is unclear for which parts of speech it should be extracted. In such situations, AmpEngine will be able to suppress the data presented in the basic container in favour of the attached one. The data from other Mansi dialects can likewise be merged and supported.

6. Conclusion

To sum it up, we believe that out multidialectal corpus is a significant contribution to the field of corpus linguistics (due to the innovative methods used to solve the problem of cross-dialectal variation), as well as to documentation of the extremely under-described Mansi language. Since the texts presented in the corpus cover the span of almost 200 years, the corpus also allows us to perform diachronic studies and research the mechanisms of Mansi dialectal divergence. Our corpus is freely accessible on the Internet (http://digital-mansi.com/corpus), which allows scholars interested in either linguistic typology or Uralic languages gain access to a valuable collection of diverse Mansi texts.

References

1. Arkhangelskiy T., Belyaev O., Vydrin A. (2012), The creation of large-scale annotated corpora of minority languages using UniParser and the EANC platform, Proceedings of COLING 2012: Posters. Mumbai: The COLING 2012 Organizing Committee, ch. 9. pp. 83–91.
2. Bakhtijarova T., Dinislamova S. (2016), A Mansi-Russian dictionary [the Upper Lozva dialect], OOO «Format», Tjumen'.
3. Balandin A., Vakhrusheva M. (1958), A Mansi-Russian dictionary [Mansijsko-russkij slovar'], Uchpedgiz, Leningrad.
4. Chernetsov V., Chernetsova A. (1936), A short Mansi-Russian dictionary [Kratkij mansijsko-russkij slovar'], Uchpedgiz, Moscow.
5. Gerstenberger C., Partanen N., Rießler M., Wilbur J. (2017), Utilizing Language Technology in the Documentation of Endangered Uralic Languages, The Northern European Journal of Language Technology 4: pp. 29-47.
6. Horváth C., Szilágyi N., Vincze V., Nagy A. (2017), Language technology resources and tools for Mansi: an overview. Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages, Saint-Petersburg.
7. Jarrar M., Habash N., Akra D., Zalmout N. (2014), Building a Corpus for Palestinian Arabic: a Preliminary Study, Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar.

[3] A container is a set of grammar data, which are transferred to the morphological analyzer to manage the annotation process. One container can include data used for annotation of texts in one language or dialect.

[4] The basic container is the one based on standard Mansi, namely [Rombandeeva 1973]; the attached container is the list of modifications that were introduced to adapt the parser for proper processing of the dialect in question. Thus, during container merging, the basic container is being modified by the set of instructions introduced by the attached container.

8. Kannisto A. (2014), A Vogul dictionary [Wogulisches Wörterbuch], Société finno-ougrienne, Helsinki.
9. Kulonen U. M. (2007), Eastern Mansi grammar and texts [Itämansin kielioppi ja tekstejä], Société Finno-Ougrienne, Helsinki.
10. Moshagen Sj., Rueter J., Pirinen T., Trosterud T., Tyers F. M.. (2014), Open-source infrastructures for collaborative work on under-resourced languages, In Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, LREC, pp. 71–77.
11. Munkácsi B., Kalman, B. (1986), A Vogul dictionary [Wogulisches Wörterbuch], Akadémiai Kiadó, Budapest.
12. Ostler N. (2008), Corpora of less studied languages, Corpus Linguistics. An International Handbook. Vol 1, pp. 457-484.
13. Rombandeeva E. I. (1973) The Mansi (Vogul) language [Mansijskij (vogul'skij) jazyk], Nauka, Moscow.
14. Rombandeeva E. I., Kuzakova E. A. (1982), A Mansi-Russian and Russian-Mansi dictionary [Slovar' mansijsko-russkij i russko-mansijskij], Prosveshenie, Leningrad.
15. Rombandeeva E. I. (2005), A Russian-Mansi dictionary [Russko-mansijskij slovar'], Mirall, Saint-Petersburg.
16. Simon E., Mus N. (2017), Languages under the influence: Building a database of Uralic languages, The 3rd International Workshop for Computational Linguistics of Uralic Languages, St. Petersburg, Russia, 23–24 January 2017, pp. 10–24.
17. Schön Zs., Wisiorek A. (2016), Ob-Ugric Database: Corpus and Lexicon Databases of Khanty and Mansi Dialects, Second International Workshop on Computational Linguistics for Uralic Languages in Szeged, Hungary.
18. Székely G. (2012). Fieldwork among Ivdel Mansi [Expedicii k ivdel'skim mansi], Issues in onomastics [Voprosy onomastiki], Vol. 1 (12), pp. 95–101.
19. Zhornik D. O., Pokrovskaya S. V. (2017), Documentation of the Upper Lozva Mansi dialect [Dokumentacija verhnelozvinskogo dialekta mansijskogo jazyka], Minor languages in big linguistics [Malye jazyki v bol'shoj lingvistike], Moscow State University, Moscow, Russia.