Towards a corpus-based grammar of Upper Lozva Mansi: diachrony and variation

Daria Zhornik, Sophie Pokrovskaya, Vladimir Plungian

Lomonosov Moscow State University
Institute of Linguistics, Russian Academy of Sciences

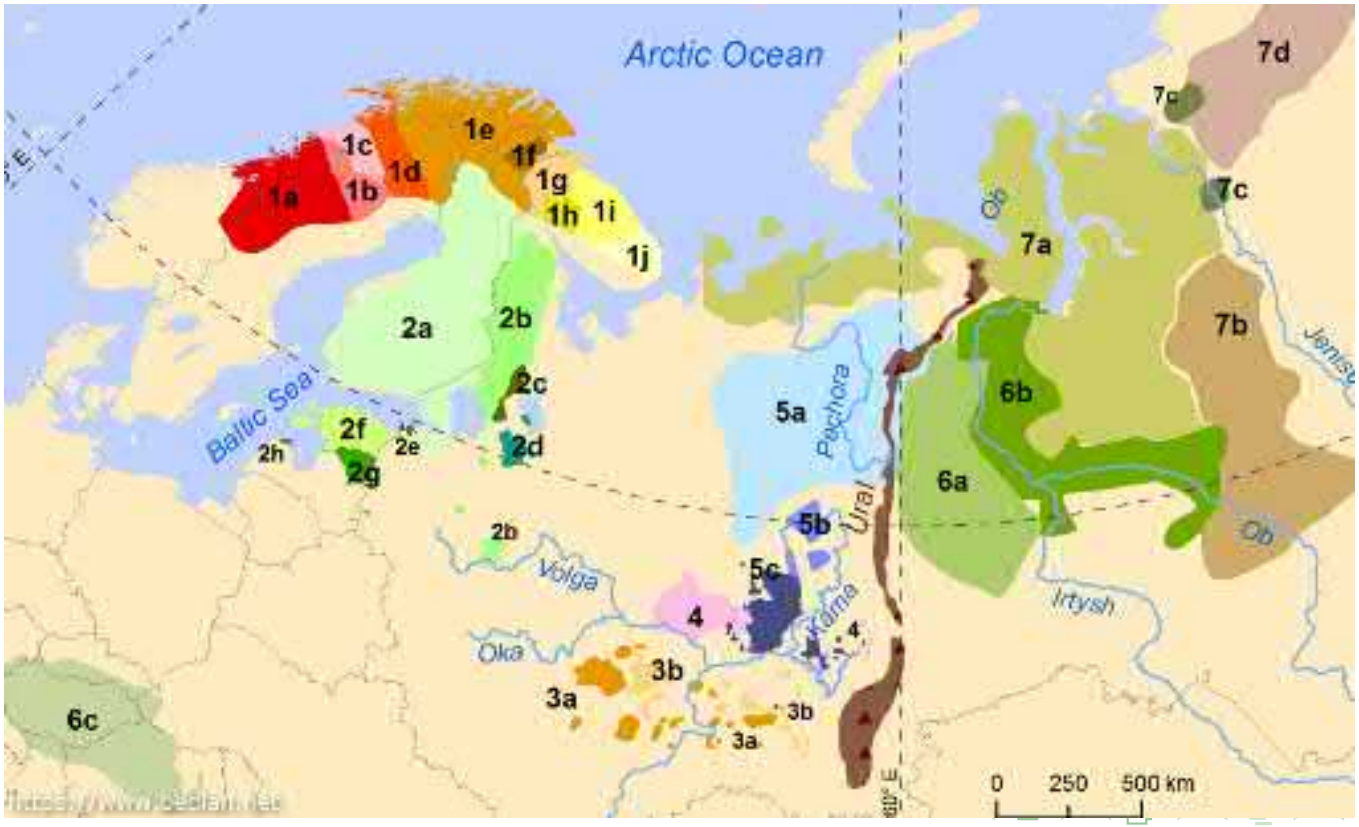
sofie.v.pokrovskaya@gmail.com
daria.zhornik@yandex.ru
plungian@gmail.com

Descriptive Grammars and Typology, Helsinki, 29.03.2019

Overview

Some background

- **Mansi** < Ob-Ugric < Uralic, EGIDS status: threatened;
- Russia, 940 speakers (2010 census) in Western Siberia.



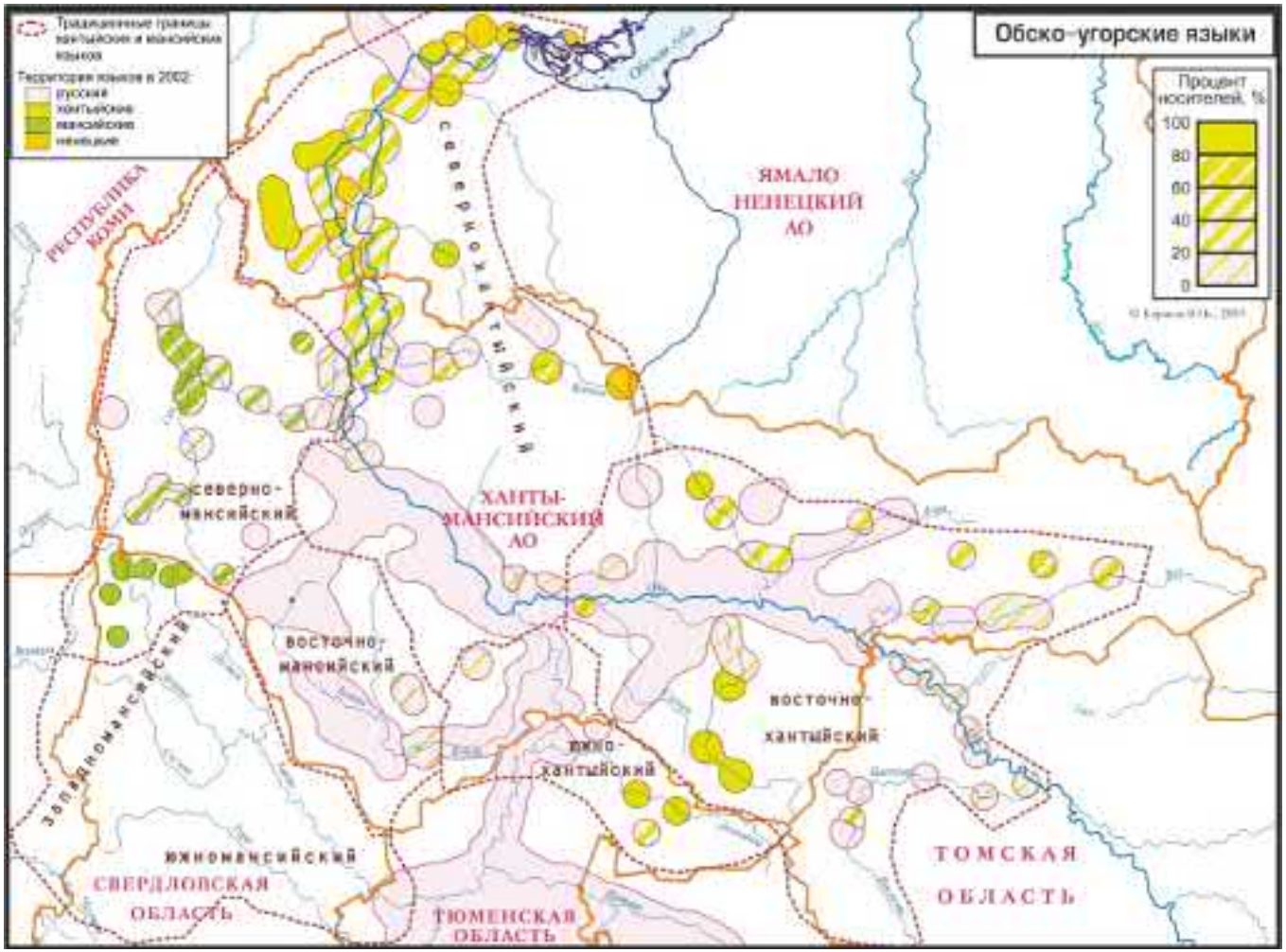
The Mansi language

- Earlier: 4 Mansi dialect branches
- Today: only the Northern group survives
- Most Northern Mansi speakers: disperse distribution in the Khanty-Mansi Autonomous Okrug
- Speakers mostly live in Russian villages/towns:
  - forced shift to Russian;
  - almost no opportunities for using Mansi;
  - all education is in Russian
- In turns out that much less people really use Mansi than the 940 mentioned in the census data.

Upper Lozva Mansi

- Upper Lozva (UL) dialect: the only well-retained Mansi variety
- Speakers live in the Sverdlovsk oblast, being geographically and administratively separated from the main mass of Mansi people
- The exact amount of speakers is unknown: 61 according to 2010 census, 108 according to some official data from 2017; the real number might be somewhere in between

Mansi dialects (map by Yuri Koryakov)



Living conditions

- Unlike most Mansi people from KhMAO, UL Mansi live in a closed community deep in the forest (150 km from the nearest small town Ivdel)
- Mansi settlements:
  - Ušma: approx. 20 speakers
  - Treskol'je (10km away): 7 permanent residents + frequently visiting relatives
  - several more distant settlements (40-80km away) containing up to 5 people each
- UL Mansi have no access to electricity or most communications (internet, mobile connection etc.)

Living conditions

Languages of interaction

- All speakers are somehow related, most UL Mansi people originate from two families: Anjamovs and Baxtijarovs.
- They use exclusively Mansi in communication with each other, and switch to Russian only to interact with outsiders.
- Members of the community differ in terms of mobility and variety of interactions:
  - some speakers travel to “civilization” frequently and have business with Russians;
  - others always stay in their home villages and may leave only on exceptional occasions (to get to the hospital, withdraw money from the bank etc.)

Education

- All UL children are fluent in Mansi and only start learning Russian at school;
- However, as they are forced to attend a boarding school (120km away from home), their use of Russian increases dramatically as soon as they begin their studies.
  - education is fully in Russian;
  - according to UL speakers, there are no competent Mansi teachers at the boarding school;
  - UL children have to communicate with everyone else in Russian
- The situation is tolerable, as UL children use Mansi to communicate with each other and during holidays they are taken back home.

Our project

- Supported by the RFBR grant №18-012-00833A *"Dynamics of phonetical and grammatical systems of Ob-Ugric languages"*
- **Team**: Daria Zhornik, Sophie Pokrovskaya (main investigators), Fyodor Sizov (software development), Vladimir Plungian (general supervision), Lev Kozlov (data processing).
- Fieldwork on the Upper Lozva dialect:
  - 3 field trips to Ushma and Treskolje in 2017-2018;
  - elicitation tasks;
  - sociolinguistic questionnaires;
  - text recordings;
  - lexicon collection;
  - experiments.

Mansi language corpus

- Multidialectal Mansi language corpus (in progress): pilot version available at `www.digital-mansi.com/corpus`
  - texts from various dialects;
  - multimedia subcorpus with fieldwork texts supplied with audio (and potentially video) files;
  - time span of almost 200 years from the beginning of the XIX century
  - big amounts of texts (although imbalanced in terms of genre, dialect etc.): newspaper subcorpus features almost 1 million tokens.

Existing descriptions

- One of our goals is of course to produce a grammar of UL Mansi after enough knowledge about the language is acquired
- This is especially important, as
  - the only existing grammars of Mansi, which are [Murphy 1968] and [Rombandeeva 1973], date almost 50 years back;
  - Murphy's description is solely based on A. Kannisto's texts gathered in 1904-1905;
  - UL data is absent from both descriptions.

Our goals in grammar writing

- A corpus-based grammar, which is
  - Ecompassing all ‘levels’ of language
  - Describing variation and language contact
  - Showing the language as a dynamic process developing from the protolanguage
  - Using empirical approach as the main principle

Corpus-based approach

- The basis for our grammar is a corpus of Mansi texts:
  - open-source data make our claims verifiable;
  - freely available audio materials allow anyone who is interested to analyze Mansi phonetics (and other areas of the language) independently
  - some phenomena can only be studied on larger pieces of discourse: prosody, information structure etc.;
  - a corpus allows us to see frequency effects;
  - individual variation is clearly reflected in the corpus, as we try to extract texts from all available speakers;
  - a well-annotated corpus allows typologists to find the desired phenomena without delving into the depths of Mansi grammar + the corpus is conveniently searchable.

Empirical approach

- ‘The empirical method is not sharply defined. It is generally characterized by the collection of a large amount of data before much speculation as to their significance, or without much idea of what to expect, and is to be contrasted with more theoretical methods in which the collection of empirical data is guided largely by preliminary theoretical exploration of what to expect’ [Brigman, Holtman 2000]

Formal approach?

- There have been numerous debates (see e.g. [Karlsson 2008], [Grewendorf 2007], [Featherston 2007]...) about empirical data and its integration in linguistic descriptions and analysis.
- One of the questions is whether it is possible to include empirical data into formal grammar and generative approach
- We believe that such models restrict the possibilities of accurately reflect some types of language phenomena
- A prominent idea in [Chomsky, Halle 1968]: *'...representing words using the phonemic rather than a phonetic representation: that is, there are some aspects of pronunciation that are redundant...'*

Importance of phonetic data analysis

- In phonological representation we may lose acoustic features and sound patterns (including variation, see [Kohlberger, last talk]) which give us information about
  - contact influence, e.g. 'non-native' sounds only appearing in loanwords etc.
  - sociolinguistic factors, for example choice of allomorph according to age group (e.g. Upper Kuskokvim differences in consonant mergers, [Kibrik, last talk])
  - idiolectal features

Importance of phonetic data analysis

- We might even lose information about linguistic phenomena from other levels of the language
- A Mansi example is phrasal stress
- Phrasal stress is marked by changing the usual quality and quantity of the stressed vowel
- Other correlates of phrasal stress (pitch, intensity) depend on a lage variety of other phenomena and cannot be reliable on their own
- The vowel under phrasal stress differs from other stressed vowels of phonologically the same quality and quantity throughout the phrase
- Thus, Mansi phrasal stress is hard to analyze in phonological terms, as it is a fine-grained phonetic detail

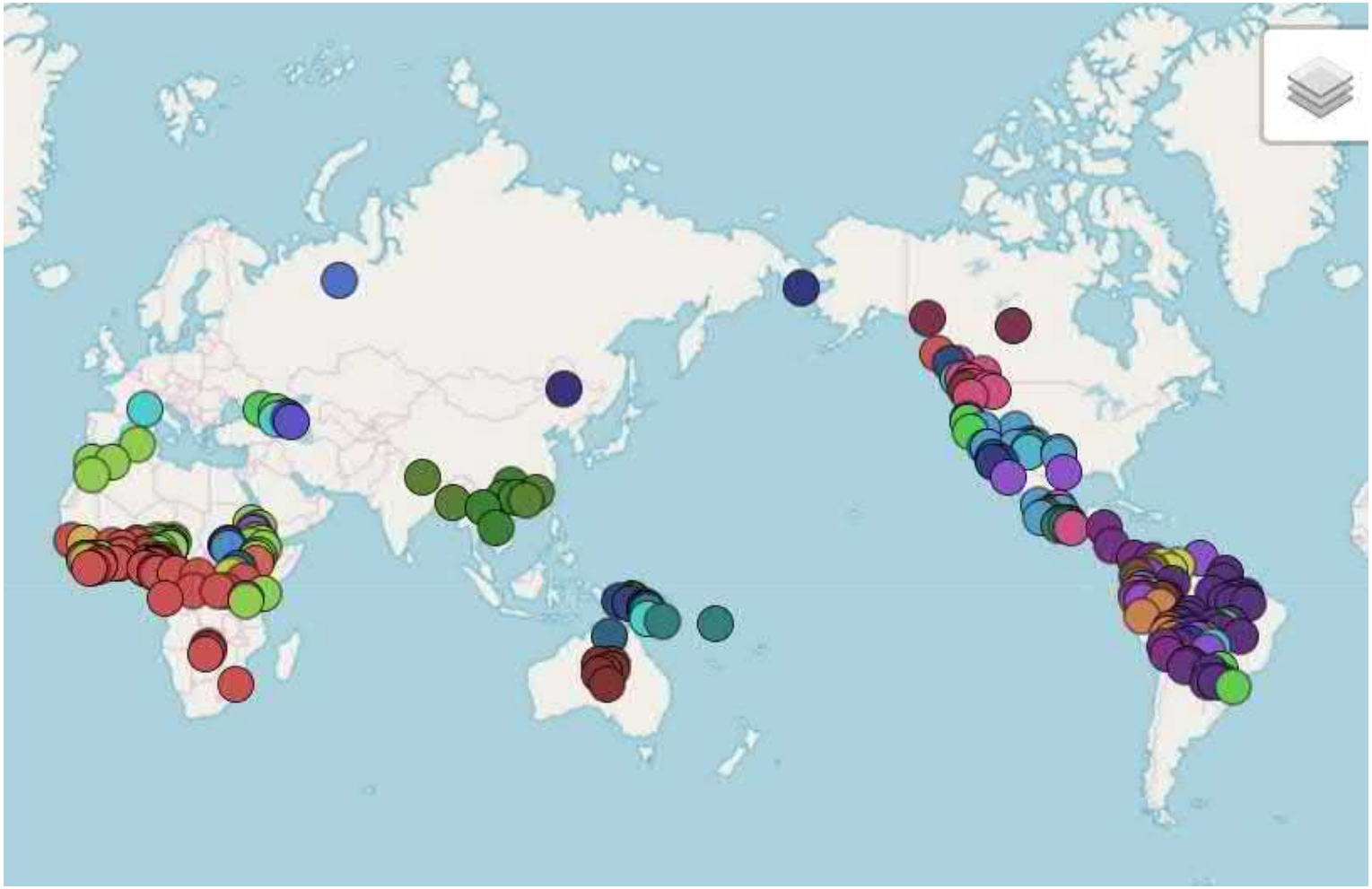
Formal approach in phonetics?

- Does this mean that formal approach completely refrains from studying phonetics?
- In [Cohn 2007]: 'To reach a fuller understanding of the workings of phonology, phonetics, the lexicon, and their interactions, we need to ... ask in an empirically-based way what is the connection between these domains of the linguistic system. What is called for are non-reductionist integrated approaches.'
- The idea of 'Translation Model' (see [Browman, Goldstein 1986]) is to model speech production based on separate articulatory features.

Gradual nature of (phonetic) phenomena

- However, separate articulatory features are not sufficient for an accurate desciption of how speech functions, they are abstract and unnatural too
- Numerous linguistic works emphasize the gradual nature of linguistic phenomena: [Plungian 2011], [Ayoun et al. 2018] to name but a few
- A Mansi example is the evolution of the complex phoneme <kw>, which manifests in different ways across speakers: different labialization degree, various types of voicing, presence or absence of the [w] segment and so on.

Random fact (Map from PHOIBLE 2.0)



Gradual nature of phonetic phenomena

- Another good example are Hill Mari ‘plosives’: depending on numerous factors, we see various degrees of voicing and different positioning on the ‘plosive-fricative’ scale, these variants are mixed with each other and illustrate a complex process in the phonological system.
- If we analyze sounds in terms of separate articulatory features we have the risk of losing valuable information on diachrony, phonology, sociolinguistics and even morphology (e.g. affixes are distinguished by phonemes which are built upon sounds) etc.
- A formal model would restrict our possibilities of reflecting natural data, leading us into the realm of abstractions.

Links between phonetics and other levels

- Phonetic data provide us with physical material and it seems logical to start defining various phenomena from the ‘most visible’ side
- In morphology, phonetic analysis gives us a basis for classification of affixes into grammatical categories, not searching for non-existent explanations of allomorphy
- In Mansi (as in numerous other languages), some aspects of verbal morphology are closely intertwined with information structure, which has its prosodic marking
- Studying the acoustic features of prosodic marking aids us in research on information structure and, in turn, even verbal morphology

A broad scope of description

- In our opinion, to provide an accurate description of the language, we have to take into account
  - all language varieties: different idiolects ("Each speaker is a dialect", [A. A. Kibrik, last talk]) and, if possible, different dialects;
  - language contact: [Forker, last talk];
  - all levels of language, ranging from phonetics to discourse structure: [Hannß, last talk], [A. A. Kibrik, last talk] etc.;
  - different diachronic stages of the language.
- All available info on these matters should be consulted;
- But it is also important to approach the language in an unbiased manner, as a *tabula rasa*, see methods proposed by A. E. Kibrik, e.g. in [Kibrik 2006]

Importance of the diachronic aspect

- ‘Language cannot be fully understood without reference to its evolution’ [Givón 2002], see also [Aikhenvald 2015] and countless other works
- A synchronic description alone is not capable of capturing all aspects of the language
- We are interested in describing the diachronic development of a language at various stages
- Data on diachronic development may also contribute to studies in language acquisition ([Givón 2015], [Bergs, Brinton 2012] etc.)

Importance of the diachronic aspect

- It might seem that a descriptive grammar does not analyze the language as a competence while describing emirical data
- However, the description of a language in its dynamic development means systematization of how various components interact, which is, in turn, an analysis
- If we observe the dynamic processes in the language, we can make predictions about its further development
- We can also see the most flexible spots in the system

Importance of the diachronic aspect

- Synchronic phenomena (especially, variation) reflect underlying diachronic processes
- A Mansi example is allophonic variation of the phoneme <ɣ>: [ɣ]/[j]
- This variation reflects the phonological merger of phonemes <ɣ> and <j>
- Velar-to-approximant developments are in general common for Ob-Ugric dialects (see e.g. [Zhivlov 2007])
- This tendency in velars may be typologically relevant as well

Interaction with native speakers

- The question of native speakers in grammar writing has provoked many discussions, see e.g. the one in [Jones, Mooney 2017];
- While a linguist can provide analysis, only the speaker can accurately process data;
- For us, the main consultant of this type is a 27-year old UL speaker Tatyana Bakhtiyarova, to whom we are extremely grateful.

Speaker-linguist workflow

- We cannot transcribe texts directly in the field due to absence of electricity and impossibility of using computers, so we split the process:
  1. the linguists record texts in the field;
  2. the speaker performs primary annotation in ELAN: segmentation into clauses, phonological transcription and translation into Russian
  3. the linguists check the annotation
  4. the linguists perform glossing
  5. the speaker consults linguists on the accuracy of glossing
  6. the speaker-linguist team discusses tricky issues
  7. the speaker-linguist team has successfully annotated data!

Conclusion

- To sum it up, our UL Mansi grammar should:
  - be corpus-based
  - be empirical
  - not be formal
  - account for variation (especially in areal Ob-Ugric context)
  - have a broad scope in terms of language levels and varieties
  - include the diachronic dimension

Thank you! Kiitos! Tack! Спасибо! Пӯмаҫипа!