

Relativization on Subjects

- Relative pronoun
- Non-reduction
- Pronoun-retention
- Gap

APmkNon
 APmk=Case
 APmk=WO
 APmk=Top

NoDWO
 ?DWO
 SV
 VS








NoPP
 NoCase
 NoPPCase
 PPAff
 AffPP
 PPCaseTon
 PPCaseInflex
 CaseMix
 &Suppl

Typological Database of the Ugric Languages

An introduction

Helsinki, 16.9.2013

Nikolett F. Gulyás, Eötvös Loránd University

- | | | | | | | | |
|---|-----------|---|----------------|---|----------|---|-------|
|  | Hungarian |  | Komi - zyrian |  | Finnish |  | Mari |
|  | Mansi |  | Komi - permyak |  | Estonian |  | Erzya |
|  | Khanty |  | Udmurt |  | Nenets | | |

Overview

- ▶ **Introduction**
- ▶ New approaches (and old questions) in Uralistics
- ▶ Linguistic typology and Uralistics
- ▶ The Uralic Typological Database Project
- ▶ Databases and other online sources
- ▶ UgTDB as a pilot project
- ▶ Summary

About our Department

- ▶ Since 1872
- ▶ Majors in Finnish, Estonian and Uralistics
- ▶ Ca. 100 students (BA, MA, PhD)
- ▶ Languages taught:
 - ▶ Finnish and Estonian
 - ▶ Veps, Karelian
 - ▶ Erzya
 - ▶ Mari
 - ▶ Komi, Permyak, Udmurt
 - ▶ Surgut-Khanty, Northern Mansi

Main research areas

- ▶ Literature, Archeology, Ethnography, History etc.
- ▶ Comparative linguistics:
 - ▶ historical comparative morphology and syntax
 - ▶ synchronic comparative morphology and syntax
 - ▶ language contacts, areal linguistics
- ▶ Sociolinguistics:
 - ▶ code-switching
 - ▶ attitudes
- ▶ Phonetics:
 - ▶ comparative research on vowel features
 - ▶ accent of non-native speakers
- ▶ Linguistic typology:
 - ▶ diachronic typology
 - ▶ TAM, alignment patterns, information structure, syntax-semantics interface
 - ▶ **databases**

Overview

- ▶ Introduction
- ▶ **New approaches (and old questions) in Uralistics**
- ▶ Linguistic typology and Uralistics
- ▶ The Uralic Typological Database Project
- ▶ Databases and other online sources
- ▶ UgTDB as a pilot project
- ▶ Summary

New approaches (and old questions) in Uralistics

- ▶ **The gap in research:**
 - ▶ early research focused on etymology
 - ▶ first descriptions: rich data without conventional descriptions, problems with transcriptions etc.
 - ▶ from '30s there were no fieldworks
 - ▶ there is a gap between the development of modern linguistic descriptions and new fieldworks carried out systematically from the '80s
- ▶ **We already have:**
 - ▶ the major FU languages investigated within different frameworks
 - ▶ new (text) corpora on minor FU languages
 - ▶ precise description on phonological and (to some extent) morphological features of minor FU languages

New approaches (and old questions) in Uralistics

- ▶ **Fields emerging in present day research:**
 - ▶ creating new text corpus (parallel and colloquial corpora)
 - ▶ systematic syntactic description
 - ▶ systematic phonetical description
 - ▶ pragmatics
 - ▶ discourse analysis
 - ▶ L2 acquisition
 - ▶ etc.

- ▶ **New(er) approaches:**
 - ▶ generative grammar
 - ▶ functional-cognitive grammar
 - ▶ construction grammar
 - ▶ variational typology
 - ▶ etc.

Overview

- ▶ Introduction
- ▶ New approaches (and old questions) in Uralistics
- ▶ **Linguistic typology and Uralistics**
- ▶ The Uralic Typological Database Project
- ▶ Databases and other online sources
- ▶ UgTDB as a pilot project
- ▶ Summary

Linguistic typology and Uralistics

- ▶ **From Uralistics to Linguistic Typology:**
 - ▶ updated and comparable data
 - ▶ detailed explanations
 - ▶ new data provided by natives (more informants)
 - ▶ possibly new questions

- ▶ **From Linguistic Typology to Uralistics:**
 - ▶ methods
 - ▶ new (-old) questions and research areas
 - ▶ the place of our language family within the world's languages

Overview

- ▶ Introduction
- ▶ New approaches (and old questions) in Uralistics
- ▶ Linguistic typology and Uralistics
- ▶ **The Uralic Typological Database Project**
- ▶ Databases and other online sources
- ▶ UgTDB as a pilot project
- ▶ Summary

The Uralic Typological Database Project

► Goals:

- an online database containing all the possible morphosyntactic parameters of a given language
- values of the parameters with examples and descriptions
- new parameters (and languages) can be added to the database

► Research questions:

- What are the typologically defined morphological and syntactic features of the analyzed Uralic languages?
- How do these languages fit into the typology of the world's languages?
- How can a database compatible with already existing digital analysis (parameters, parameter values, combinative contrastive mechanisms) be created in theory and practice for researchers to analyze relevant relationships between languages?

UTDB

- ▶ Once again, the reasons:
 - ▶ intragenetic typology analyzes related languages in terms of their shared features and differences
 - ▶ Uralic languages have not been analyzed from this point of view before
- ▶ History of the project:
 - ▶ 2005: CIFU, Joshkar-Ola: the aim has emerged
 - ▶ 2008: Vienna: conference, pilot projects on different typological questions
 - ▶ 2010: CIFU, Piliscsaba: symposium dedicated to the typological description of the Finno-Ugric languages
 - ▶ 2013- The Typological Database of the Ugric Languages (pilot project)

UTDB

- ▶ The database:
 - ▶ it provides all the basic material for further typological collection
 - ▶ it will be accessible to not only Finno-Ugric linguists but also researchers of linguistic typology
 - ▶ it will offer both complete typological analysis of the Uralic languages in terms of morphology and syntax and make it possible to do comparative analysis with the rest of the world's languages
 - ▶ the parameters and programming framework of the database can offer an example and starting point for similar future projects

UTDB

- ▶ **Some examples:**
 - ▶ word order
 - ▶ marking of semantic roles
 - ▶ possession
 - ▶ case marking
 - ▶ alignment patterns
 - ▶ argument structure
 - ▶ negation
 - ▶ the structure of the NP, VP
 - ▶ person marking
 - ▶ numerals

Overview

- ▶ Introduction
- ▶ New approaches (and old questions) in Uralistics
- ▶ Linguistic typology and Uralistics
- ▶ The Uralic Typological Database Project
- ▶ **Databases and other online sources**
- ▶ UgTDB as a pilot project
- ▶ Summary

Databases and other online sources

- ▶ **Typological databases:**
 - ▶ **The World Atlas of Language Structures**
 - ▶ **Matthew Dryer's Typological Database**
 - ▶ **Surrey Morphology Group**

WALS

- ▶ the most complex typological database of the world's languages
- ▶ data presented from 2678 languages
- ▶ main categories:
 - ▶ phonology
 - ▶ nominal categories
 - ▶ verbal categories
 - ▶ nominal syntax
 - ▶ word order
 - ▶ simple clauses
 - ▶ complex sentences
 - ▶ lexicon, sign languages, other

WALS

- ▶ <http://wals.info/>
- ▶ Uralic languages:
 - ▶ 27 languages
 - ▶ traditional classification, but: Mari and the Mordvinic languages are separated
- ▶ Parameters: ca. 160
 - ▶ the most for Hu: 140
 - ▶ the fewest for Izhor and Liv: 1
- ▶ Notes:
 - ▶ Yazva-Komi is classified as a Finnic language
 - ▶ Khanty, Mansi, Karelian, Estonian presented without separation
 - ▶ only 3 Saami languages
 - ▶ some values are questionable

Matthew Dryer's Typological Database

- ▶ only word order parameters for areal patterns
- ▶ 34 different parameters
- ▶ 225 languages including 34 Uralic languages
- ▶ not easy to search data and compare the results
- ▶ search by language (family) is not possible
- ▶ the Ob-Ugric languages have only a few values

- ▶ <http://linguistics.buffalo.edu/people/faculty/dryer/dryer/database>

Surrey Morphology Group

- ▶ investigation of grammatical categories with the use of explicit formal and statistical frameworks
- ▶ different databases on canonical typology
- ▶ a very good source for particular questions
- ▶ Hungarian is included but not in all cases
- ▶ Ob-Ugric languages are neglected

- ▶ <http://www.surrey.ac.uk/englishandlanguages/research/smg/webresources/index.htm>

Databases and other online sources

- ▶ Languages and language families:
 - ▶ Ethnologue
- ▶ Online etymological dictionary:
 - ▶ Uralonet
- ▶ On Ob-Ugric languages:
 - ▶ Ob-Babel project
- ▶ Morphologic analyzers
- ▶ the Hungarian National Corpus & the Mazsola:
 - ▶ HNC
 - ▶ Mazsola

Ethnologue

- ▶ General information about the world's languages:
 - ▶ ISO
 - ▶ Alternate Names
 - ▶ Population
 - ▶ Location
 - ▶ Language Maps
 - ▶ Language Status
 - ▶ Classification
 - ▶ Dialects
 - ▶ Typology (only basic word order)
 - ▶ Language Development
 - ▶ Language Resources
 - ▶ Writing
 - ▶ **Place in Language Cloud**
- ▶ Notes:
 - ▶ there is no Ugric subgroup
 - ▶ the dialectal grouping at least for Khanty is controversial
 - ▶ <http://www.ethnologue.com/>

Uralonet

- ▶ Digitalized version of the Uralisches Etymologisches Wörterbuch(UE W
- ▶ some updated etymologies
- ▶ new comments can be added
- ▶ meanings are provided in Hungarian, German and English
- ▶ search not only by languages but by semantic fields as well
- ▶ some exercises according to the comparative method can be found
- ▶ <http://www.uralonet.nytud.hu/>

Ob-Babel project

- ▶ a collection of former data about the Ob-Ugric languages
- ▶ <http://www.babel.gwi.uni-muenchen.de/>
- ▶ **Content:**
 - ▶ text corpora in four different dialects of Khanty and Mansi
 - ▶ innovative e-grammars
 - ▶ e-dictionaries
 - ▶ an e-library
 - ▶ ethnographic and folklore material
- ▶ **Notes:**
 - ▶ morphologic analyzer is useful but there are some problems with the tool in use
 - ▶ typological data is not provided
 - ▶ NB: different tasks as in the UTDB

Morphologic analyzers

- ▶ Analyzers for Nganasan, Permic and Ob-Ugric languages
- ▶ Text corpora is included
- ▶ Structural analysis
- ▶ Task: complex morphologic analyzer
- ▶ The tool works well with its own corpus but cannot deal with other transcriptions (at least in the case of Khanty)
- ▶ <http://www.morphologic.hu/urali/>

The Hungarian National Corpus & the Mazsola

▶ HNC:

- ▶ the largest digital corpus for Hungarian
- ▶ based on written sources
- ▶ different dialects, materials from the whole Carpathian Basin
- ▶ different genres: press, literature, science, official, personal
- ▶ more than 187 million tokens

▶ Mazsola:

- ▶ on the basis of the HNC
- ▶ for the research on argument structures: verbal prefixes, adjectives etc.
- ▶ textual frequency

Overview

- ▶ Introduction
- ▶ New approaches (and old questions) in Uralistics
- ▶ Linguistic typology and Uralistics
- ▶ The Uralic Typological Database Project
- ▶ Databases and other online sources
- ▶ **UgTDB as a pilot project**
- ▶ Summary

UgTDB as a pilot project

- ▶ objective: an online database for the Ugric languages
- ▶ content:
 - ▶ parameters
 - ▶ description of the parameters
 - ▶ languages: Northern Mansi, Surgut-Khanty, Synja-Khanty and Hungarian
 - ▶ description of the languages and different subgroups
 - ▶ features
 - ▶ description of the features
 - ▶ values
 - ▶ description of the values
 - ▶ glossed examples for a given value in all languages
 - ▶ citations
 - ▶ search tool (dealing with 5 different parameters at the same time)
 - ▶ the database works in English, in Russian and in Hungarian

Overview

- ▶ Introduction
- ▶ New approaches (and old questions) in Uralistics
- ▶ Linguistic typology and Uralistics
- ▶ The Uralic Typological Database Project
- ▶ Databases and other online sources
- ▶ UgTDB as a pilot project
- ▶ **Summary**

Summary

- ▶ New approaches and research questions in recent times
- ▶ Cross-linguistically relevant questions have remained unanswered in Uralistics
- ▶ Some endangered languages are well documented, minor FU languages are not
- ▶ A database can provide a possible solution for this problem
- ▶ UTDB > UgTDB
- ▶ There are a lot of databases to follow, none of which is completely suitable for our goals > UgTDB

Next time











- ▶ Our project
- ▶ Tasks
- ▶ Target languages
- ▶ Sources
- ▶ Methods
- ▶ Parameters
- ▶ Values
- ▶ Recent results
- ▶ Questions
- ▶ More questions

Number of Cases



KIITOS! KÖ SZÖ NÖM THANK Y

nikolett.fgulyas@gmail.com

-  Hungarian
-  Komi - zyrian
-  Finnish
-  Mari
-  Mansi
-  Komi - permyak
-  Estonian
-  Erzya
-  32 Khanty
-  Udmurt