Zsófia Schön

OUDB Yugan Khanty corpus

Second workshop on Uralic prosody

RIL HAS

Budapest, 28.09.2017

Outline

1. Project Background
2. Fieldwork Recordings
3. Corpus

zsofia.schoen@gmail.com

1. Project Background








zsofia.schoen@gmail.com

1.1 Project Background

OUDB

- Ob-Ugric database:
  analyzed text corpora and dictionaries
  for less described Ob-Ugric dialects
- 1st July 2014 – 30th June 2017
- 1–5 team members for the Yugan Khanty team

http://oudb.gwi.uni-muenchen.de/

zsofia.schoen@gmail.com

1.2 Project Background

OUDB

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

OB-UGRIC DATABASE
ANALYSED TEXT CORPORA AND DICTIONARIES FOR LESS DESCRIBED OB-UGRIC DIALECTS

VIENNA MUNICH

Index | Sitemap | LMU-Homepage | Login

ENGLISH | РУССКИЙ | DEUTSCH

- INTRODUCTION
- ABOUT OUL
- ABOUT OUDB
- ENCYCLOPEDIC SECTION
- LINGUISTIC SECTION
  - Text Corpus
    - All Corpora
    - All Mansi Corpora
    - All Khanty Corpora
    - Northern Mansi
    - Eastern Mansi
    - Northern Vagilsk Mansi
    - Pelym Mansi
    - Middle Lozva Mansi
    - Lower Lozva Mansi
    - Kazym Khanty
    - Surgut Khanty
    - Yugan Khanty (1991)
    - Yugan Khanty (2010–)
  - Concordance
  - Transcription, Transliteration and Orthography
  - e-Dictionaries
  - e-Grammar
  - Download
- LINKS

The **OUDB Yugan Khanty corpus (YK)** contains texts from an unpublished collection by Zsófia Schön, partly together with Lyudmila Nikolaevna Kayukova. The transcription depicts spoken language by means of a broad phonematization using IPA characters. Only words or, respectively, sentences from native speakers of Khanty are glossed and included in the lexicon.

Yugan Khanty (2(▼ | optional dialect ▼ | all genres ▼ | glossed only | audio only | all translation ▼ | Search

sort by original title ▼ | annotated only

Original_Title	Dialect	English_Title	ID	Genre_Form	Genre_Content	glossed	annot.	Audio	Translation
e:j βo:jyəliyən o:piseyən (ENK)	yugan khanty (YK)		1533	prose (pro)	Tales (tal)	glossed	–	Audio	
e:mp tju:n to:nəp i:ki (TMK)	yugan khanty (YK)	Old-dog-backtendoned-man (TMK)	1514	prose (pro)	Ethnographic Texts (eth)	glossed	–	Audio	en dt ru
e:mp tju:n to:nəp i:mi (AIK)	yugan khanty (YK)	Old-dog-backtendoned-woman (AIK)	1515	prose (pro)	Ethnographic Texts (eth)	glossed	annot.	Audio	en dt ru
e:mp tju:n to:nəp i:mi (SPK)	yugan khanty (YK)	Old-dog-backtendoned-woman (SPK)	1518	prose (pro)	Ethnographic Texts (eth)	glossed	–	Audio	en dt ru
i:ttən əntə nɛripti (TMK)	yugan khanty (YK)	Why one shouldn't misbehave at night (TMK)	1507	mixed (mix)	Ethnographic Texts (eth)	glossed	–	Audio	en dt ru
i:ttən əntə nɛripti (AJM)	yugan khanty (YK)	Why one shouldn't misbehave at night (AJM)	1523	prose (pro)	Tales (tal)	glossed	–	Audio	en dt ru
i:ttən əntə nɛripti (AIK)	yugan khanty (YK)		1538	mixed (mix)	Tales (tal)	glossed	–	Audio	ru
i:ttən əntə nɛripti (JFP)	yugan khanty (YK)	Why one shouldn't misbehave at night (JFP)	1542	mixed (mix)	Tales (tal)	glossed	–	Audio	en dt ru
oŋq tu:təŋkəliyən o:piseyən 2. (TMK)	yugan khanty (YK)	The nuthatch and his older sister 2. (TMK)	1479	prose (pro)	Tales (tal)	glossed	–	Audio	en dt ru
oŋq tu:təŋkəliyən o:piseyən 1. (TMK)	yugan khanty (YK)		1596	mixed (mix)	Tales (tal)	glossed	–	Audio	
n'l mp l	yugan khanty	The strange little	1526	prose (pro)	Tales (tal)	glossed	–	Audio	en dt ru

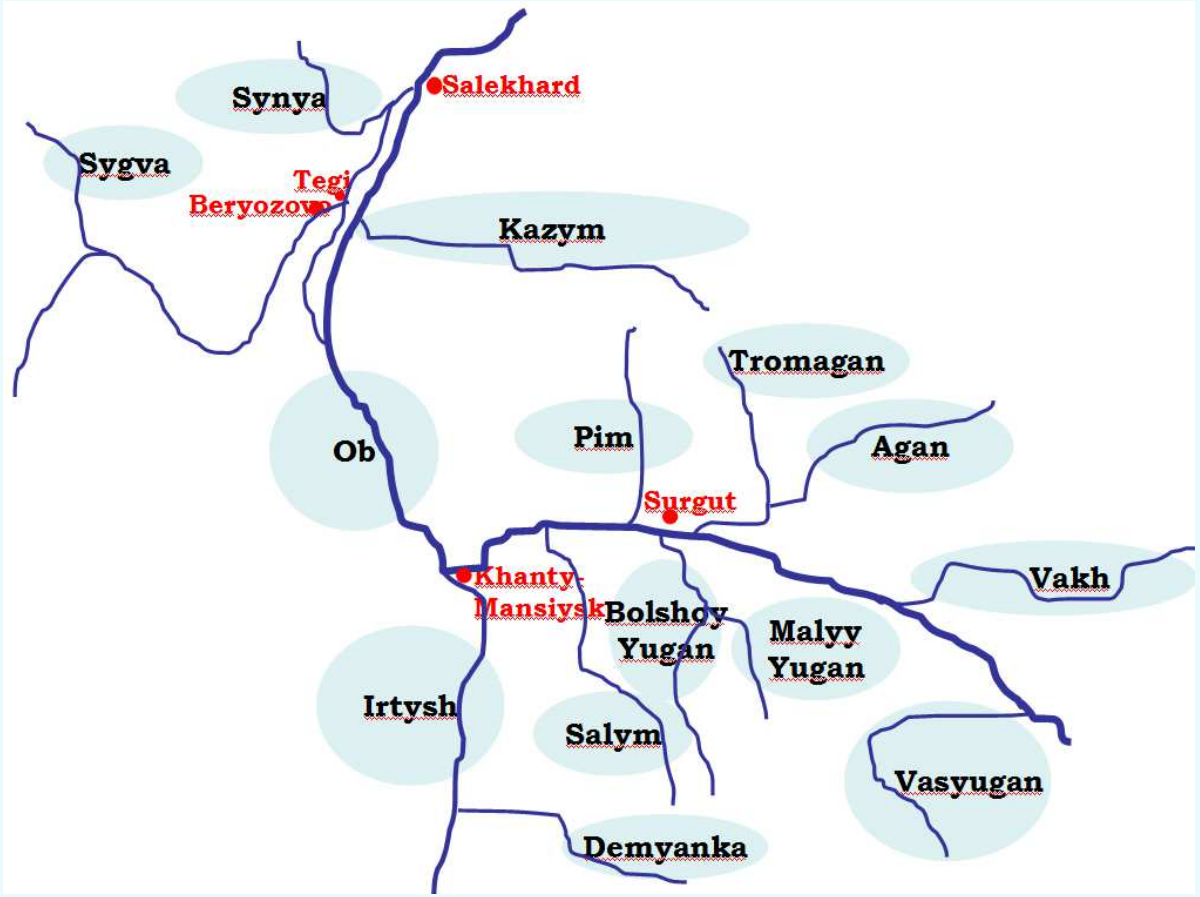
= 30 texts

zsofia.schoen@gmail.com

1.3 Project Background

Yugan Khanty

- spoken in North-West Siberia, along the rivers Bolshoy Yugan, Malyy Yugan, Salym and Demyanka



zsofia.schoen@gmail.com

1.4 Project Background

Yugan Khanty

- ca. 900 native speakers
- Eastern Dialect Group > Surgut Khanty
- only "south" subdialect of Surgut Khanty
- phonotactic: V1 + /ɣ/ + V2 > V1ː

(1) *puːɣəɬ ~ puːːɬ*
'village'

(2) *nʉŋ kiːnʲtʲeː*
nʉŋ kiːnʲtʲe-ɣe
2sg than-2sg
'than you'

zsofia.schoen@gmail.com

2. Fieldwork Recordings








zsofia.schoen@gmail.com

2.1 Fieldwork Recordings

Collectors:
- Ludmila Nikolaevna Kayukova
- Zsófia Schön

Collection:
- 29 texts, 1 h 16 min 15 sec

Genre:
- tales
- ethnographic texts
- personal accounts

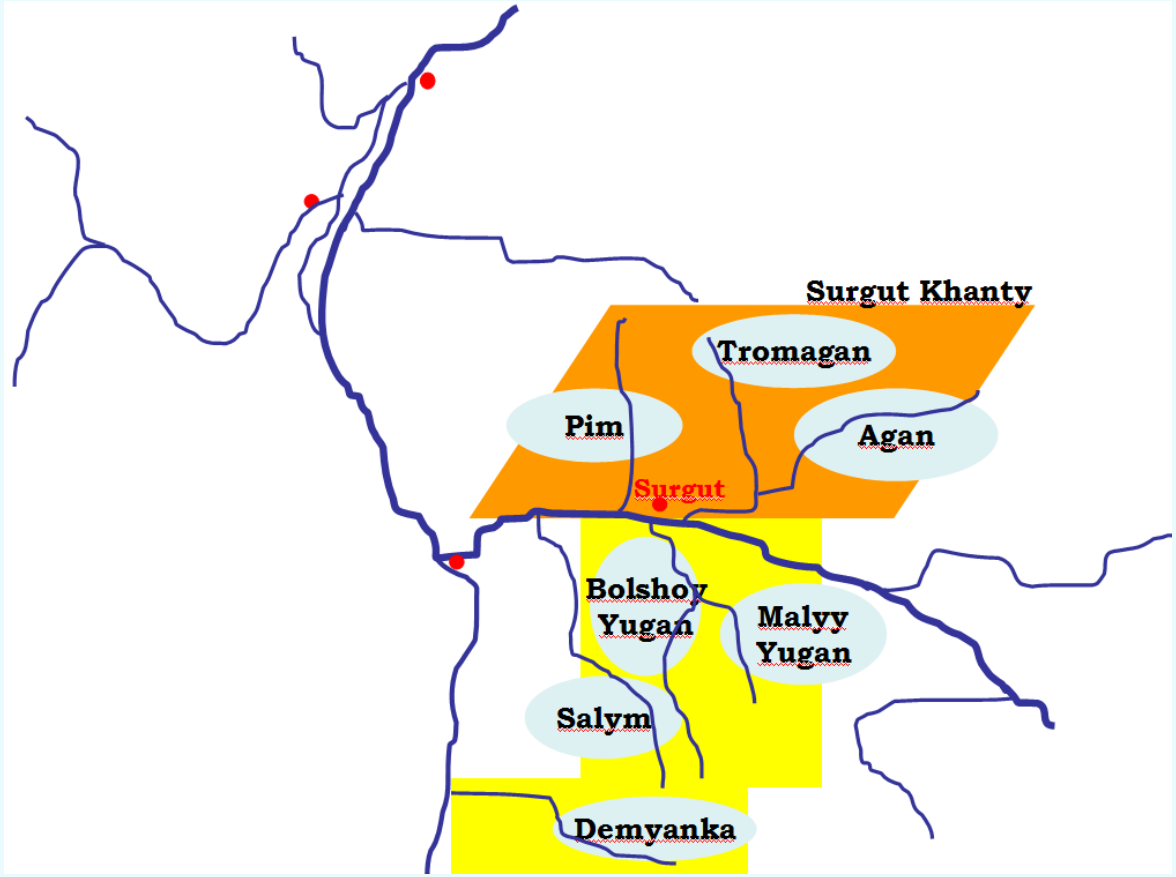
Form:
- prose
- dialogues
- prose + dialogues

zsofia.schoen@gmail.com

2.2 Fieldwork Recordings

Informants:

- 13 – 12 female and 1 male speaker
- Bolshoy Yugan
- Malyy Yugan
- Salym
- Demyanka



zsofia.schoen@gmail.com

2.3 Fieldwork Recordings

Equipment:

- 2010–2012 Olympus LS10
- 2015–2017 Zoom H5
- WAV 44.1kHz/16bit
- stereo
- built-in microphones
- fixed position in front of the speaker

zsofia.schoen@gmail.com

2.4 Fieldwork Recordings

Circumstances: • no laboratory cirsumstances






zsofia.schoen@gmail.com

3. Corpus








zsofia.schoen@gmail.com

3.1 Corpus

Contains:
- 29 texts, 1 h 16 min 15 sec

Transcription:
- IPA
- wide phonemic
- non-standardized, non-normatized
- spoken language
- speech disfluency
- elison, epenthesis

zsofia.schoen@gmail.com

3.2 Corpus

Workflow:

- segmentation and transcription in ELAN (29)
- check, correction and on-the-fly Russian translation with L.N. Kayukova
- glossing in FLEx (27)
- translation into English (22), German (22) and Russian (23)
- import of sound file, ELAN-Export and FLEx-Export in the database/user interface

zsofia.schoen@gmail.com

3.3 Corpus

Data access: http://oudb.gwi.uni-muenchen.de/

- Metadata

Text + Metadata | Translation | Audio + Metadata | Glossed Text

Original Title	Dialect	Informant	Genre Form	Genre Content	ID	glossed	Audio
iːttən ənta nɛripti (TMK)	yugan khanty (YK)	Kel'mina, Tat'yana Mikhaylovna (Kaymysova)	mixed (mix)	Ethnographic Texts (eth)	1507	by Schön, Zsófia	Audio

Text Source	Editor	Collector
First publication Zsófia Schön (2016)	Schön, Zsófia; Kayukova, Lyudmila Nikolaevna	Schön, Zsófia (ZS)

English Translation	German Translation	Russian Translation	Hungarian Translation
"Why one shouldn't misbehave at night (TMK)" by Eckmann, Stefanie; Schön, Zsófia	"Wieso man nachts nicht unartig sein soll (TMK)" by Eckmann, Stefanie; Schön, Zsófia	"Почему ночью нельзя баловаться (TMK)" by Kayukova, Ludmila Nikolaevna; Snigirev, Yury; Schön, Zsófia	-

Citation

Schön, Zsófia 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1507. Ed. by Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1507 (Accessed on 2017-09-27)

Audio Metadata:

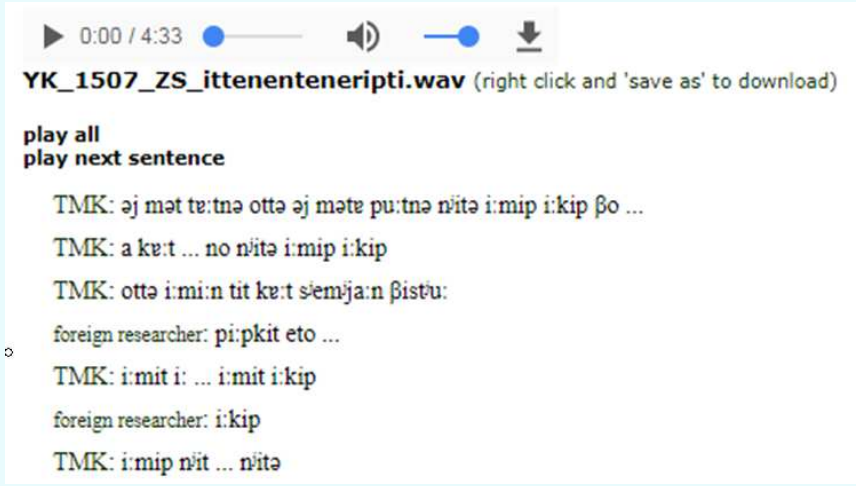
Informant	Year of Birth (Informant)	Subdialect (Informant)	Collector	Date of Recording	Place of Recording	recorded by
Kel'mina, Tat'yana Mikhaylovna (Kaymysova)	1966	Yugan Khanty	Schön, Zsófia (ZS)	2012	Yurty Achimovy II	Schön, Zsófia

zsofia.schoen@gmail.com

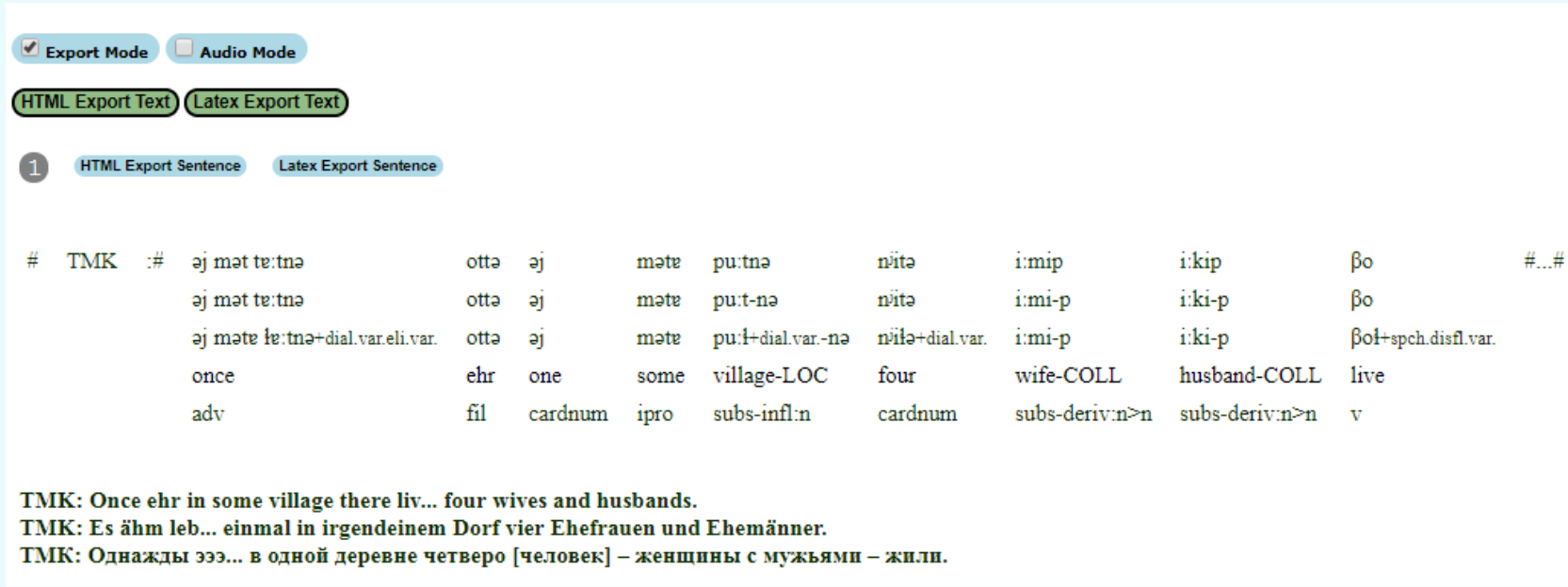
3.4 Corpus

Data access: http://oudb.gwi.uni-muenchen.de/

- Audio + Metadata



- Glossed Text

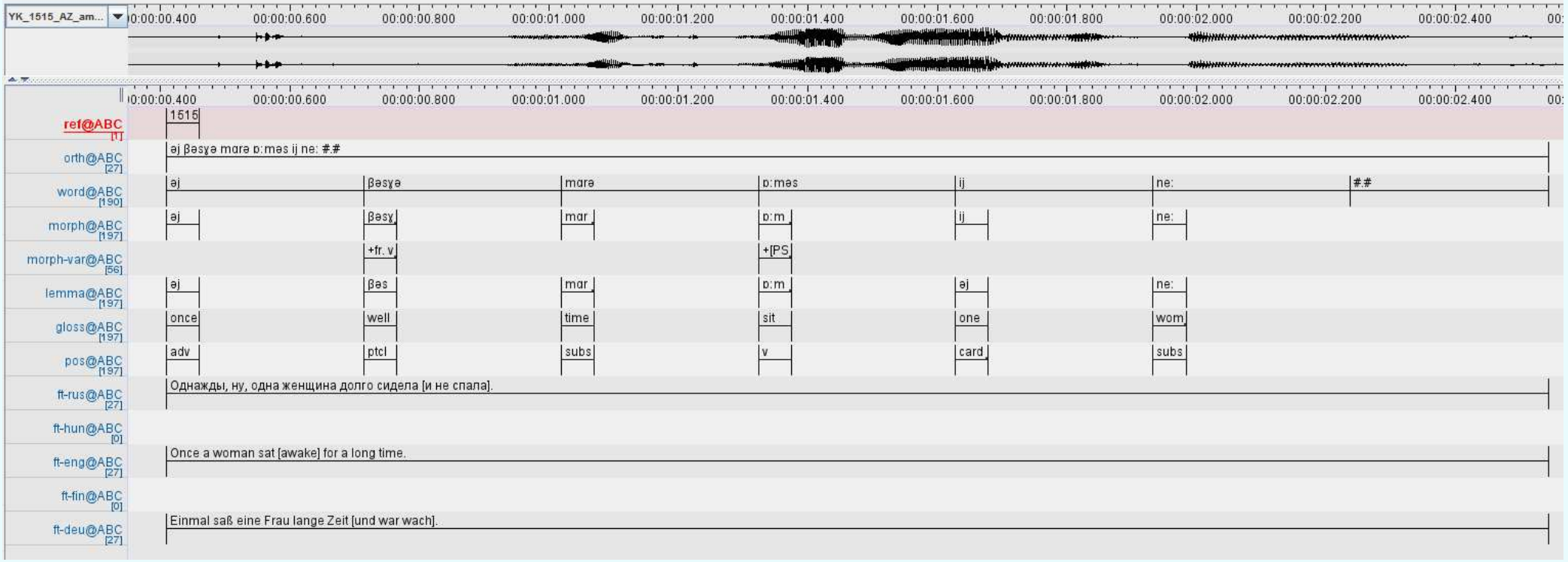


#	TMK	:#										#...#
			əj mət tɐ:tnə	ottə	əj	mətɐ	pu:tnə	nʲitə	i:mip	i:kip	βo	
			əj mət tɐ:tnə	ottə	əj	mətɐ	pu:t-nə	nʲitə	i:mi-p	i:ki-p	βo	
			əj mətɐ łɐ:tnə+dial.var.eli.var.	ottə	əj	mətɐ	pu:ɬ+dial.var.-nə	nʲiłə+dial.var.	i:mi-p	i:ki-p	βoɬ+spch.disfl.var.	
			once	ehr	one	some	village-LOC	four	wife-COLL	husband-COLL	live	
			adv	fil	cardnum	ipro	subs-infl:n	cardnum	subs-deriv:n>n	subs-deriv:n>n	v	

TMK: Once ehr in some village there liv... four wives and husbands.
TMK: Es ähm leb... einmal in irgendeinem Dorf vier Ehefrauen und Ehemänner.
TMK: Однажды эээ... в одной деревне четверо [человек] – женщины с мужьями – жили.

3.5 Corpus

Data access:
- The Language Archive
- 17 of the 29 Yugan Khanty texts
- WAV, ELAN-, FLEx-Data combined
- small metadata



zsofia.schoen@gmail.com

3.6 Corpus

Futur plans:

- orthographic transcription, literarization
  > book of Yugan Khanty tales
- expansion of the corpus
- search for new funding and projects

zsofia.schoen@gmail.com

Bibliography

Чепреги, Марта. 2017. Сургутский диалект хантыйского языка. Ханты-Мансийск.

Gugán, Katalin–Schön, Zsófia. 2018. 9.5 East Khanty. In: Bakró-Nagy, Marianne – Laakso, Johanna – Skribnik, Elena (eds) Oxford Guide to the Uralic Languages.

Скрибник, Елена К.–Шён, Жофия–Янда, Гвен Ева–Визиорек, Аксель–Снигирёв, Юрий–Экманн, Штефани. 2017. Обско-Угорская база данных:текстовые корпусы и словари Обско-Угорских диалектов". Ханты-Мансийск.

Wisiorek, Axel–Schön, Zsófia. 2016. Obugric Database: Corpus and Lexicon Databases of Khanty and Mansi Dialects. In: Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages. Szeged.

Wisiorek, Axel–Schön, Zsófia. 2017. Obugric Database: Corpus and Lexicon Databases of Khanty and Mansi Dialects. Acta Linguistica Academica 64,3. Budapest.

zsofia.schoen@gmail.com

pe:sipe βołitəɣ









http://oudb.gwi.uni-muenchen.de/

zsofia.schoen@gmail.com