Anja Behnke – Josefina Budzisch

Selkup Language Corpus

Working Papers in Corpus Linguistics
and Digital Technologies:
Analyses and Methodology
Vol. 6.

Anja Behnke – Josefina Budzisch

Selkup Language Corpus

Working Papers in Corpus Linguistics and Digital Technologies:
Analyses and Methodology
Vol. 6.

Szeged – Hamburg
2021

Working Papers in Corpus Linguistics and Digital Technologies: Analyses and methodology
Vol. 6

WPCL issues do not appear according to strict schedule.

Vol. 6 (2021)

Published by
University of Szeged, Department of Finno-Ugric Studies
Egyetem utca 2. 6722 Szeged
Universität Hamburg, Zentrum für Sprachkorpora
Max-Brauer-Allee 60 22765 Hamburg

Published 2021

DOI 10.14232/wpcl.2021.6

ISSN 2677-0857
ISBN 978-963-306-838-0 (pdf)

Contents

Tables, figures and maps

1.Introduction

This paper documents the project: “Syntactic Description of the Southern and Central Selkup Dialects: A Corpus-Based Investigation”, which was carried out between 2015 and 2018 at the University of Hamburg. The project was funded by the German Research Foundation (DFG). The main goal of the project was the creation of a digital language corpus of Selkup. In addition to the originally planned texts from Central and Southern Selkup dialects, a number of Northern Selkup texts were added in the course of the project. The corpus, therefore, reflects the great dialectal diversity of Selkup.

The paper is structured as follows: Section 2 describes the project objectives and the tasks that were carried out during the course of the project. In section 3, a short overview of Selkup is presented, giving some remarks about the areal distribution as well as the linguistic status of Selkup. In section 4, metadata about the corpus are introduced; here information about archiving and conventions throughout the corpus are described. Section 5 deals with the structure of the corpus and gives a detailed analysis of the transcription and annotation of the data. In section 6, a list of research based on the corpus is presented, section 7 lists the text sources for the corpus, and in section 8 references are given. In the appendix, the used characters, as well as labels for glosses and categories, can be found.

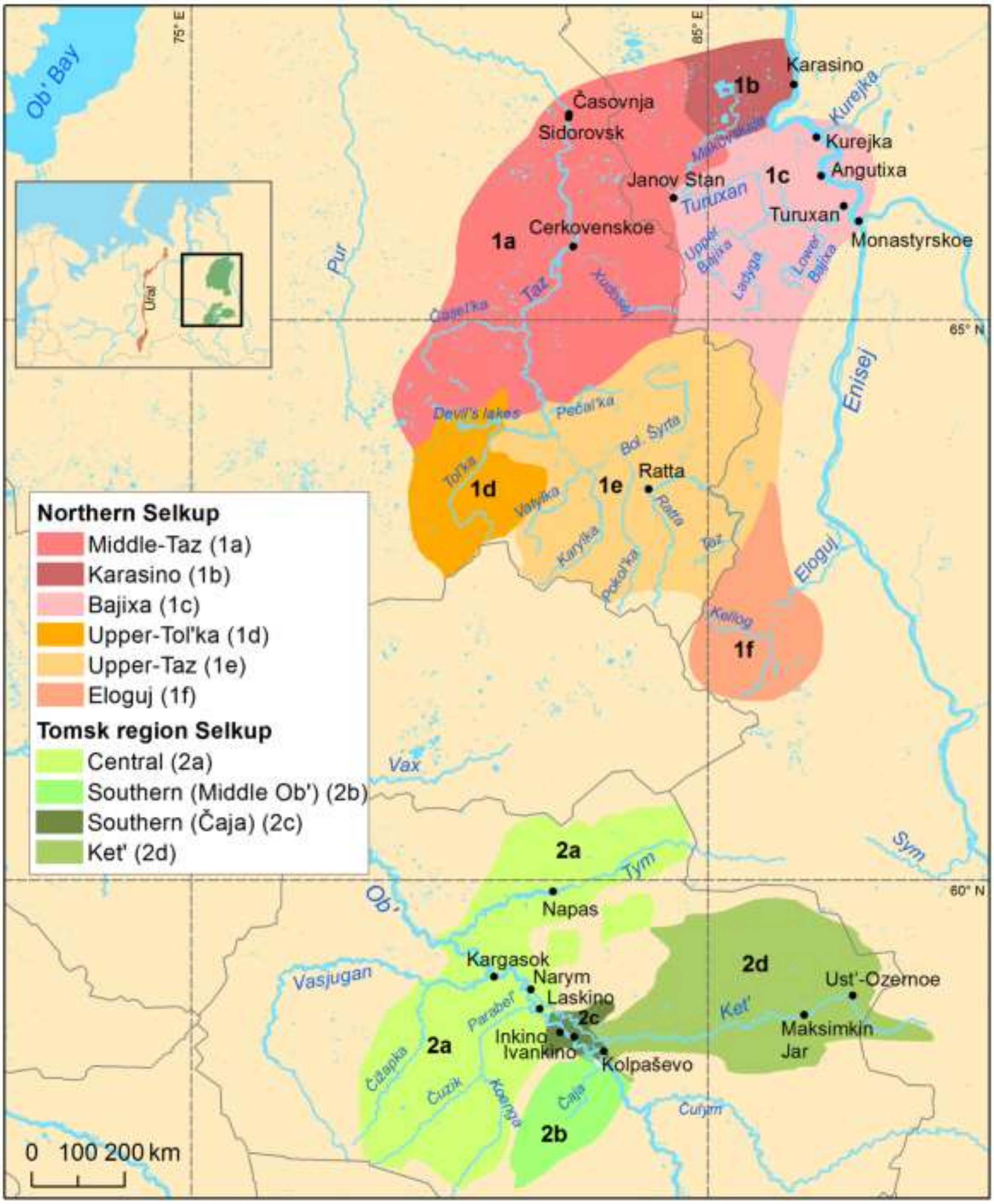
2.Project goals

The main objective of the project was to create a digital corpus of language using already published Selkup texts. The focus was placed on Central and Southern dialects, as these have so far only been described sparsely. The corpus contains all primary data and the associated metadata, which are archived together. There are two types of metadata: a) the personal data of the native speakers (some sources contain only minimal information such as name, place of residence, and date of birth; in the case of Kuz'mina's publications, the metadata in the publication is kept to a minimum, but they are verifiable and expandable with the help of the field research notes); b) information about the text itself is included, the following data points are consistently given for each individual text: speaker, the dialectal affiliation of the speaker, researcher who recorded the text, date, and place of recording. A pdf file of the publication is also provided in the corpus for all texts. In summary, the project objectives were as follows:

1. Digitizing published Selkup texts from different genres
2. Transcribing and (syntactic) annotating the texts
3. Creation of metadata
4. Compilation of a digital text corpus

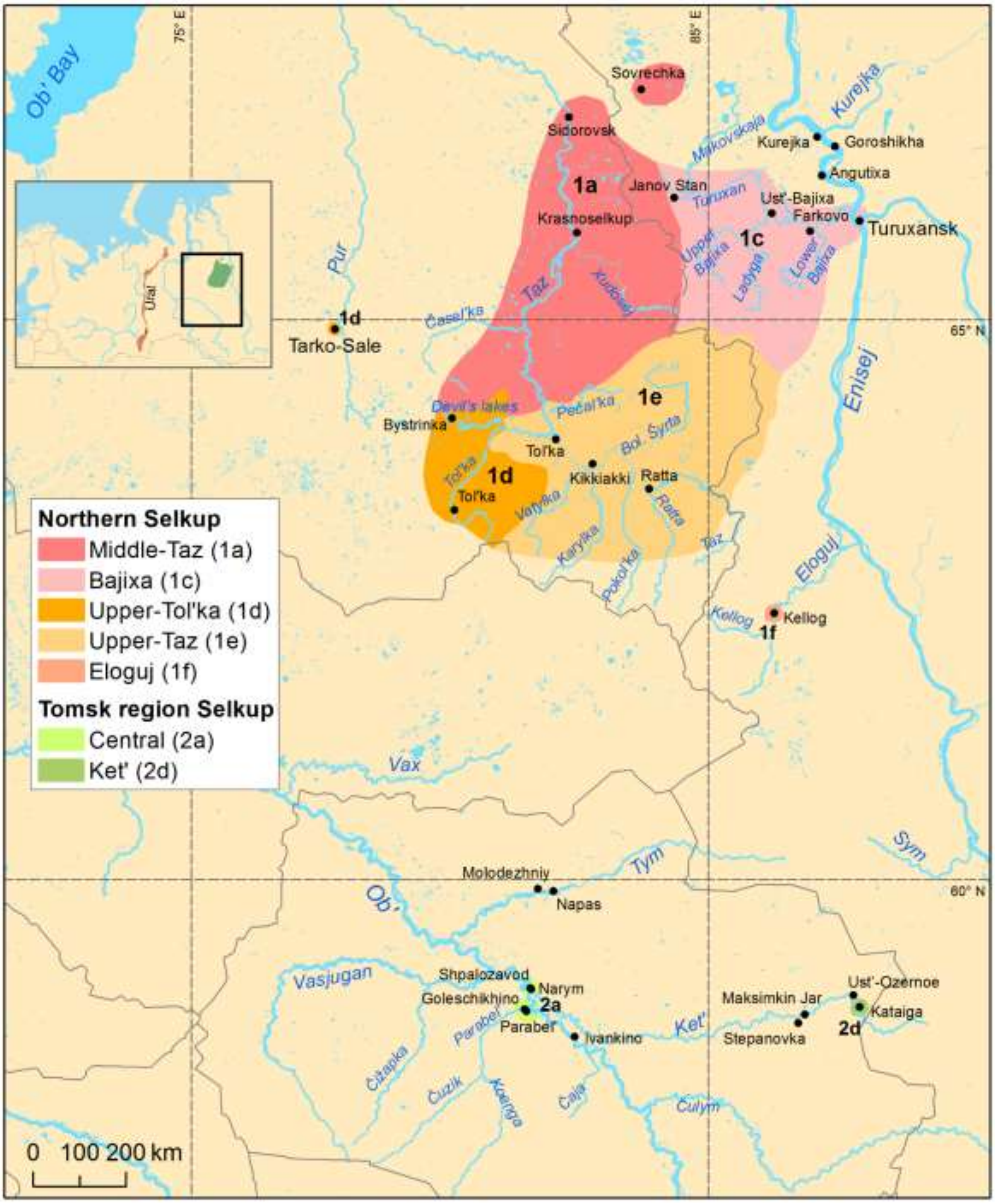
3.Selkup

Selkup, which was formerly known as Ostyak Samoyed, belongs to the South Samoyedic branch of the Uralic language family. The Selkups are widely scattered in Siberia between the rivers Ob and Yenisei. They can be divided into two larger groups, a part of the Selkups lives in the more northern areas of Siberia (along the rivers Taz, Turukhan, and the Yenisei), a second group lives further south, in the vicinity of the Ob river and its tributaries. Map 1 shows the original residential areas of the Selkups, not dialects or dialectal groups, map 2 shows the current residential areas.



Map 1 *Original Selkup settlements*[1]

[1] Source: Rantanen, Timo, Vesakoski, Outi, Ylikoski, Jussi, & Tolvanen, Harri. (2021). Geographical database of the Uralic languages (Version v1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4784188



Map 2 *Current Selkup settlements*[2]

According to the last Russian census (2010) only 3,649 Selkups still live in Russia, 1,023 people have indicated that they are speakers of Selkup. However, the distribution of the speakers is not even: The results of recent field research show that most speakers speak a variety of the northern dialect, while the Southern and Central Selkup dialects are critically endangered. According to current estimates, there are only about 5–10 active (older) speakers in the latter two dialect groups. One of the main

[2] Source: Rantanen, Timo, Vesakoski, Outi, Ylikoski, Jussi, & Tolvanen, Harri. (2021). Geographical database of the Uralic languages (Version v1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4784188

reasons why the vast majority of the Selkup speakers still alive speak a variety of Northern Selkup is the Russian expansion, which began earlier in the southern areas and was much more intense in its effects. The Selkups native to the northern areas were, therefore, able to preserve their traditional way of life and thus also their language longer than the Selkups of the southern areas.

There has been no direct contact between the two Selkup groups for a long time, which led to the fact that the languages developed independent and mutual understanding is hardly possible today. This means that the two groups are distinctly different both in terms of culture and language.

There is still no consensus among researchers about the division of the Selkup dialects into dialectal groups. Depending on the researcher, one speaks of three (e.g. Gluškov et al. 2011), four or more dialectal groups (e.g. Alatalo 1994, 2004, Helimski 1998) – Helimski, for example, regards the Ket dialect as an independent dialectal group with its own various subdialects.

As part of the project, the dialects were divided into three larger dialectal groups (cf. Gluškov et al. 2011): Northern Selkup, Central Selkup, and Southern Selkup, each of which is divided into further subdialects, sometimes with a further subdivision (e.g. Taz and Ket’). From an ethnographic point of view, on the other hand, only two groups can be distinguished, the Northern Selkups and the non-Northern Selkups (cf. Sokolova 1970: 145). Table 1 shows the structure of the dialects as used in the presented project.

Table 1 *Selkup dialects*

Northern Selkup	**Central Selkup**	**Southern Selkup**
Taz	Vakh	Middle Ob
Tolka (Laryak)	Vasyugan	Chaya
Karasino	Tym	Upper Ob
Turuchan	Narym	Ket’
Baikha		Chulym (†)
Eloguy		

At the end of the 19th century, the collection of language samples of Selkup began, but these were often only published in text collections much later: Castrén & Lehtisalo (1960), collected 1845–1849; Grigorovsky (first published in 1879, republished in Hajdú 1973 and Katz 1988). Donner's text from the Tym dialect was first published in 1968 by Hajdú and later by Katz (1975). In the 1930s, Prokofʻjev and his wife Prokof’jeva began to study the Samoyedic languages intensively. The results of their work can be found in numerous publications, including a grammar of Selkup (Prokofʻjev 1935) or in school books (Prokofʻjev & Prokofʻjeva 1940; Prokofjeva 1953).

With regard to the Southern Selkup dialects, the so-called ‘Tomsk School’ is relevant. In numerous publications (e.g. Dul'zon 1966a, b, 1971, Kuz'mina 1967, 1968, 1974) texts in the Southern or Central Selkup dialects were published that stem from field research carried out by Andreas Dul'zon and his students from the 1960s onwards. Some text collections were also published in the series “Skazki narodov sibirskogo Severa” in the 1980s. Different phonetic transcriptions make editing the texts difficult. Numerous texts have remained unpublished for a long time, and their publication has only begun in recent years (e.g. in Tučkova 2004 or Tučkova & Helimski 2010). These text publications are more consistent and reliable in terms of the quality of the transcription.

The beginnings of the grammatical description of Selkup can be located in the 19th century, when Castrén traveled to Siberia and published a grammar (1854), in which, in addition to the other Samoyed languages, he also deals with Selkup. The next grammatical description of the Selkup language came almost 100 years later: As already mentioned, in the 1930s Prokof‘jev dealt intensively with the Selkup language. His works (1931, 1935, and 1937) still provide a good starting point for exploring the Selkup language today, although his works tend to be short and based on the Northern dialects. Following Prokof‘jev, Kuznecova et al. (1980) published a modern, descriptive grammar almost 50 years later. This work was based on the field research materials collected in the 1970s. Unfortunately, to this day, this grammar remains the only full grammar that describes several aspects of language (phonetics, phonology, morphology, and syntax). The vast majority of grammatical treatises only contain phonological and morphological descriptions, most of which, like the work of Kuznecova et al., are based on the Northern dialects of Selkup.

The syntactic description of the Samoyed languages of Tereščenko (1973) is the only work that deals exclusively with the topic of syntax. The author deals with the simple sentences, repeatedly giving examples of Selkup, but here also mainly refers to Northern Selkup. Complex sentences are missing, as is the description of the noun phrase structure or the predication types.

So far, only a few studies have dealt with the grammar of the Southern and/or Central dialects: Kuz’mina (1974) and Bekker et al. (1995a, 1995b) describe in their grammars mainly the morphological properties of the Southern and Central Selkup dialects, a syntactic and also phonological investigation is completely absent here. The grammars also include the grammatical description of the Tym dialect (Katz 1975). According to today's view, this is a corpus-based description: The author evaluated the materials Kai Donner collected in the 1910s and compiled a short grammar based on this data. In addition to the word comments, this is limited to a very brief phonetic description. Helimski (1983) wrote a grammatical outline of the Southern Selkup Chaya dialect based on Grigorovsky's Selkup texts published in 1879.

In addition, some smaller essays deal with special phenomena of Selkup syntax: Alitkina (1983) in her four-page article non-verbal predicates, concentrating only on attribution, she goes on other types (such as belonging (proper inclusion) or equation), not a. Čeremisina & Martynova (1991) describe the syntactic functions of the Southern Selkup verb.

Although many of the statements made by Kuznecova et al. (1980) about the Northern dialects can very likely be transferred to the Southern and Central Selkup, there are considerable differences between the dialectal groups in some areas, which must be investigated more closely.

4.The corpus

The following chapter describes the content of the text corpus. This includes archives, researchers, texts, speakers, and date of recording.

The *Selkup language corpus* is based on written text, published in various sources beforehand. It contains texts from all three dialectal groups. The corpus contains 144 glossed and annotated texts from 48 speakers, 9,156 utterances with 55,839 tokens can be found in the corpus.

Table 2 Corpus data

	speakers[3]	texts	utterances	tokens
Northern Selkup	14	31	1,710	10,017
Central Selkup	15	48	3,459	22,131
Southern Selkup	26	69	4,359	24,874
Mixed dialects	1	4	1,737	12,047
total	48	152		

For all texts the original Selkup text is given as well as a translation to English, furthermore most texts are also translated to Russian, German and some to Hungarian.

SIL Fieldworks Explorer (*FLEx*)[4] is used to gloss the text's morphology. Afterwards the texts are exported to *EXMARaLDA*, this is carried out by Alexandr Archipov and Beáta Wagner-Nagy. In *EXMARaLDA Partitur Editor*[5] annotations for syntactic functions (SyF), semantic roles (SeR) and information status (IST) are added as well as additional annotation for some texts. The data about the texts and the metadata about the speakers are managed with *EXMARaLDA Corpus Manager* (Coma)[6].

4.1. Citation

Budzisch, Josefina – Anja Harder – Beáta Wagner-Nagy 2019. Selkup Language Corpus (SLC). Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0.0. Publication date 2019-02-08. http://hdl.handle.net/11022/0000-0007-D009-4 .

All the authors have equally contributed to the creation of the corpus and are listed here in the alphabetical order.

4.2. Abbreviations of researcher

In the data about the texts in Coma, the main researchers adding to the corpus are marked by their abbreviations, given here in alphabetical order:

BJ: Budzisch, Josefina
HA: Harder[7], Anja
WNB: Wagner-Nagy, Beáta

[3] Northern Selkup, 4 Southern Selkup and 4 mixed texts are from unknown speakers, which here are counted as one speaker.

[4] http://software.sil.org/fieldworks/support/using-sendreceive/flex-bridge/

[5] http://exmaralda.org/en/partitur-editor-en/

[6] http://exmaralda.org/en/corpus-manager-en/

[7] Anja Harder is now called Anja Behnke

4.3. Archiving

The transcriptions and metadata of the corpus are stored in EXMARaLDa format. The archiving and publication are taken care of by the *Hamburg Centre for Language Corpora* (HZSK).

4.4 Access rights

The corpus is available with a HZSK-RES-access. Therefore requires both using a valid account from a research institution to authenticate the end-user and sending a separate application to the rights holder for authorization, possibly including a research plan with the resource.[8]

4.5 Metadata in the corpus

The metadata of the corpus is provided in *EXMARaLDA Corpus Manager* (*Coma*), here each text has an individual name and is linked to its speaker. The metadata for the texts contain basic metadata as the place and date of recording as well as information about the researchers involved with the annotation of this text.

4.5.1 Naming conventions

The communications (texts) are all named the following way: the abbreviation of the speaker is given (first letter of the first name, the patronymic and the last name), followed by the year of recording, a short title and the abbreviation of the genre. Unknown speakers are abbreviated with NN and unknown years with XX. For example:

Name:	ChDN_1983_GirlAndIce_flk
Speaker code:	ChDN
Year of recording:	1983
Short title:	GirlAndIce
Genre:	folklore

In the corpus, four genres can be found:

a) Folklore texts (flk)

b) Narrative texts (nar): stories about everyday life or biographies

c) Songs (song)

d) Translations (trans): translations from Russian to Selkup

4.5.2. Text metadata

Name: The name of the text, see 2.1.

Genre: The genre of the text (flk, nar, song or trans)

Recorded by: The researcher who recorded the text

Date of recording: The date of the recording (if known)

Dialect group: Information about the dialect group (Northern, Central, Southern)

Dialect: Information about the dialects (see Table 1 above)

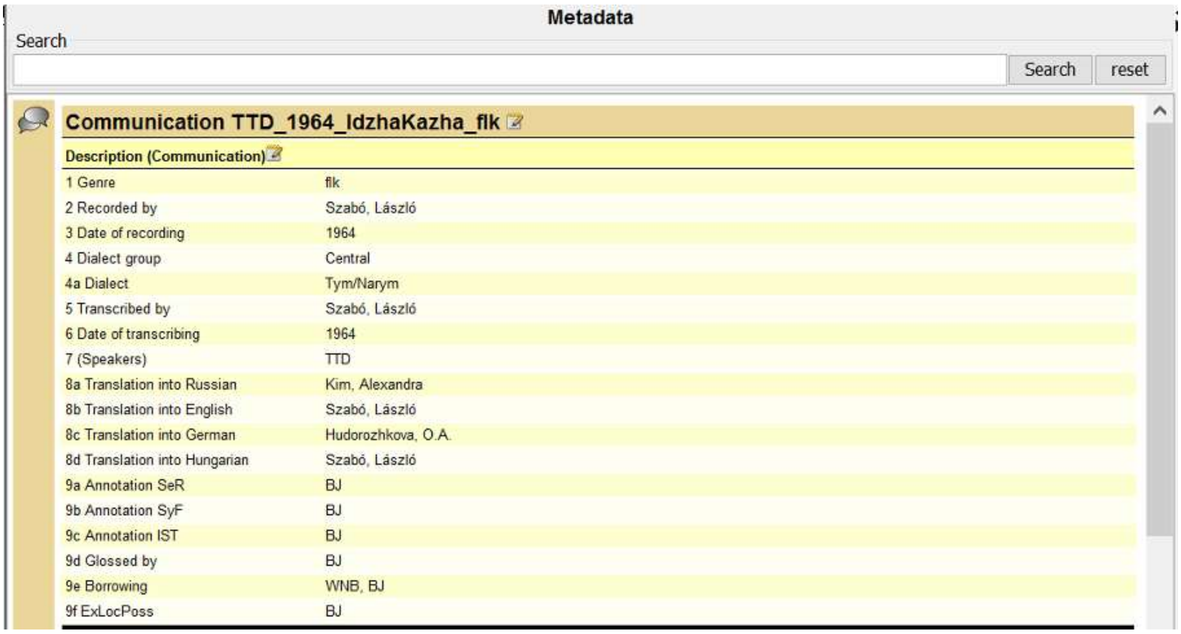
Subdialect: Information about the subdialects

Transcribed by: The researcher who transcribed the text

[8] https://corpora.uni-hamburg.de/hzsk/en/corpus-enquiries-licenses, last access: 10.08.2021

Date of transcribing: The date of the transcribing, if known
Date of translation: The date of translation (for trans), if known
Speaker: Abbreviation of the speaker
Original speaker: Abbreviation of the speaker of the original (for trans)
Translation into Russian: The researcher who translated the text. [Here given is the original translation if available. Texts without Russian translation are mostly not translated into Russian.]
Translation into Selkup: The speaker who translated the text (for trans)
Translation into English: The researcher who translated the text
Translation into German: The researcher who translated the text
Translation into Hungarian: The researcher who translated the text. [Here given is the translation in the original source if available. Texts without Hungarian translation are not translated into Hungarian.]
Glossed by: The name of the researcher, who glossed the text
Annotation SeR: The name of the annotator for semantic roles
Annotation SyF: The name of the annotator for syntactic function
Annotation IST: The name of the annotator for information status
Annotation Borrowing: The name of the annotator of borrowed elements
Annotation ExLocPoss: The name of the annotator of existential/locative/possessive sentences
Annotation CVB: The name of the annotator of converbal constructions

Figure 1 *Screenshot of text metadata (annotation data)*



Communication TTD_1964_IdzhaKazha_flk

Description (Communication)	
1 Genre	flk
2 Recorded by	Szabó, László
3 Date of recording	1964
4 Dialect group	Central
4a Dialect	Tym/Narym
5 Transcribed by	Szabó, László
6 Date of transcribing	1964
7 (Speakers)	TTD
8a Translation into Russian	Kim, Alexandra
8b Translation into English	Szabó, László
8c Translation into German	Hudorozhkova, O.A.
8d Translation into Hungarian	Szabó, László
9a Annotation SeR	BJ
9b Annotation SyF	BJ
9c Annotation IST	BJ
9d Glossed by	BJ
9e Borrowing	WNB, BJ
9f ExLocPoss	BJ

Additionally, there are given the following information:
Location:
City: the place where the text has been recorded (if known)
Country: the country where the texts was recorded (normally Russia)
LanguageCode: The language code of the text: sel – Selkup
Setting:

Archive: Information about the archive in which the text can be found, if it is known

Original text: Information about the source of the original text (for trans)

Published in: Information about previous publications. Here are all publications given in which the text was ever published.

Russian source: Information about the Russian source for texts based on Russian sources

Transcriptions: The basic and segmented transcriptions are added here.

Files: e.g. copies of archive materials or publications. Files (pdfs) are named the following way:

a) **Publication**: Author_Year_ShortTitel_Genre_Pages

The year is referring to the year of publication, the short title is the same as for the text it is belonging to, from-to page numbers are indicated by <->, if the text is on separate pages, the numbers are divided by <_>.

Example: Kuzmina_1967_Mammoth_flk_320_328 (publication of the text KFN_1967_Mammoth_flk)

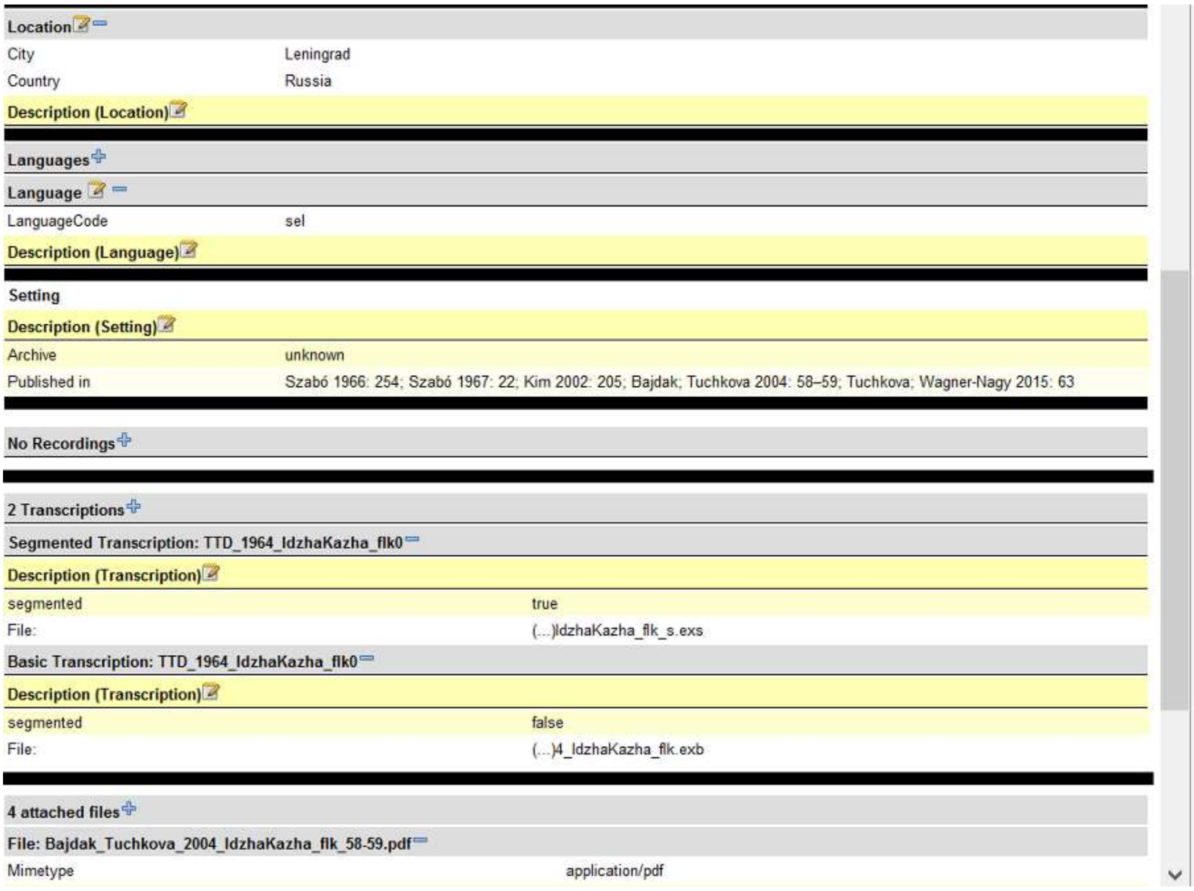
b) **Archive materials** : here are given materials from two archives:

Kuzmina archive in Hamburg: Speaker_Year_Titel_Genre_Vol_Nt_Pages

Dulzon archive in Tomsk: Speaker_Year_Titel_Genre_Vol_Pages

Example (Kuzmina archive): KFN_1967_Mammoth_flk_Vol6_Nt4_75-76 (archive material of the text KFN_1967_Mammoth_flk)

Figure 2 *Screenshot of text metadata (storing details)*



Location

City Leningrad

Country Russia

Description (Location)

Languages

Language

LanguageCode sel

Description (Language)

Setting

Description (Setting)

Archive unknown

Published in Szabó 1966: 254; Szabó 1967: 22; Kim 2002: 205; Bajdak; Tuchkova 2004: 58–59; Tuchkova; Wagner-Nagy 2015: 63

No Recordings

2 Transcriptions

Segmented Transcription: TTD_1964_IdzhaKazha_flk0

Description (Transcription)

segmented true

File: (...)IdzhaKazha_flk_s.exs

Basic Transcription: TTD_1964_IdzhaKazha_flk0

Description (Transcription)

segmented false

File: (...)4_IdzhaKazha_flk.exb

4 attached files

File: Bajdak_Tuchkova_2004_IdzhaKazha_flk_58-59.pdf

Mimetype application/pdf

4.5.3. Speaker metadata

Metadata related to the speakers include in all cases biographical information and the linguistic biography of the speaker. Further relevant data will also be included whenever it is available.

The following data is given (if known):

Description of speaker: The name of the speaker.

→ Given are: Family name, patronymic, given name

Education: Information about the education and occupation of the speaker.

→ Given are: Education, Higher education, Occupation (if it known)

Informant of: The researcher the speaker worked with.

Ethnicity: Background information about the ethnicity of the speaker and its relatives.

→ Given are: Ethnicity, Ethnicity of mother, Name of mother, Ethnicity of father, Name of father, Ethnicity of husband/wife, Name of husband/wife, Ethnicity of grandparents

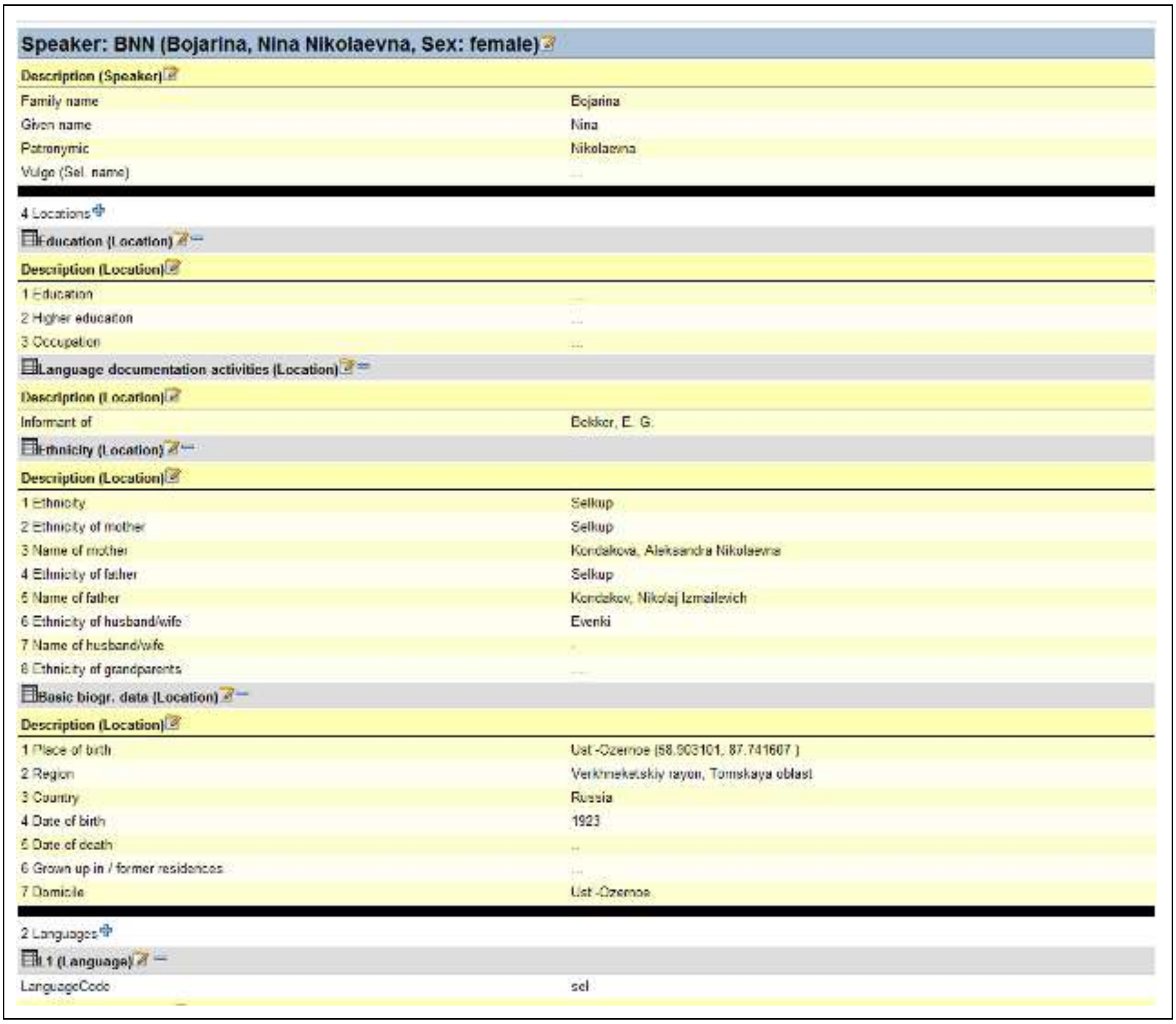
Basic biographical data: Information about the past and current places of residence and basic vital statistics; the domicile is always the current or last (in case of death) place of residence.

→ Given are: Place of birth, Region, Country, Data of birth, Data of death, Grown up in /former residences, Domicile

Languages: The speaker's languages, all speakers speak Selkup (sel) and Russian (rus).

→ Given are: L1, L2

Figure 3 *Screenshot of speaker metadata*



Speaker: BNN (Bojarina, Nina Nikolaevna, Sex: female)	
Description (Speaker)	
Family name	Bojarina
Given name	Nina
Patronymic	Nikolaevna
Vulgo (Sel. name)	...
4 Locations	
Education (Location)	
Description (Location)	
1 Education	...
2 Higher education	...
3 Occupation	...
Language documentation activities (Location)	
Description (Location)	
Informant of	Bekker, E. G.
Ethnicity (Location)	
Description (Location)	
1 Ethnicity	Selkup
2 Ethnicity of mother	Selkup
3 Name of mother	Kondakova, Aleksandra Nikolaevna
4 Ethnicity of father	Selkup
5 Name of father	Kondakov, Nikolaj Izmailevich
6 Ethnicity of husband/wife	Evenki
7 Name of husband/wife	-
8 Ethnicity of grandparents	...
Basic biogr. data (Location)	
Description (Location)	
1 Place of birth	Ust-Ozernoe (58.903101, 87.741607)
2 Region	Verkhneketskiy rayon, Tomskaya oblast
3 Country	Russia
4 Date of birth	1923
5 Date of death	..
6 Grown up in / former residences	...
7 Domicile	Ust-Ozernoe
2 Languages	
1 (Language)	
LanguageCode	sel

4.6 Archives

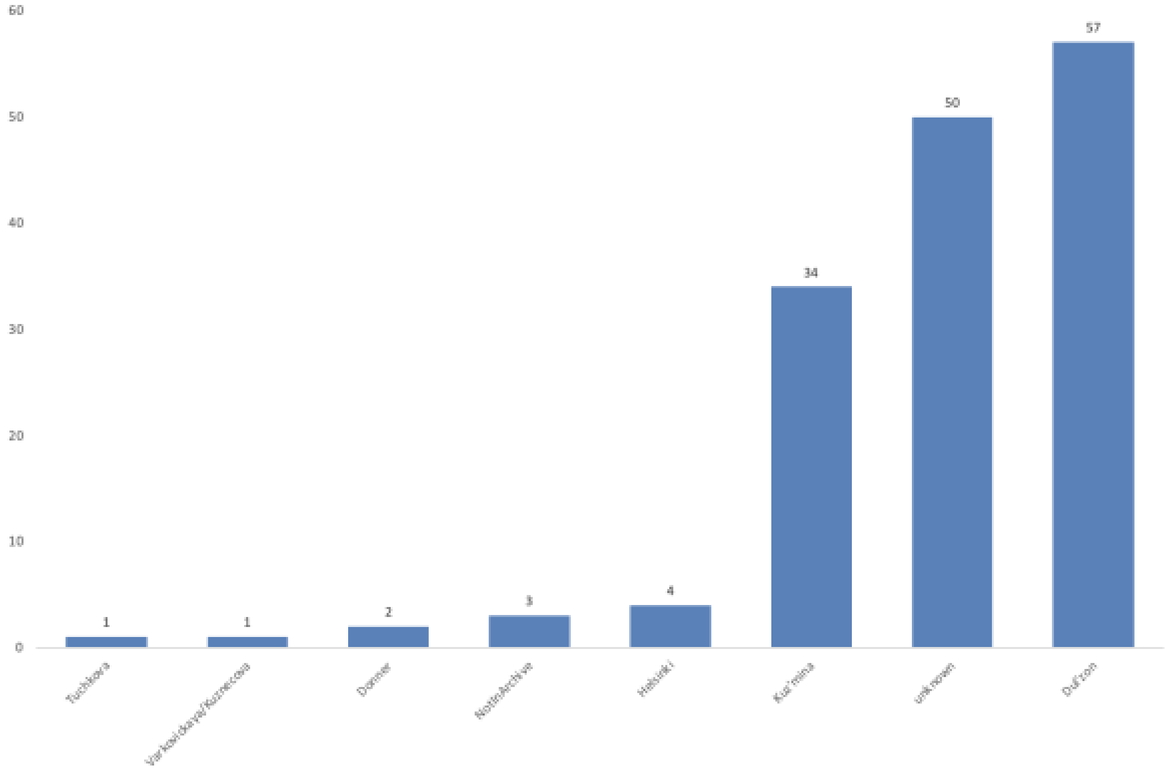
The corpus is based on published texts, the originals of which, however, are stored in different archives. Figure 5 shows the number of corpus texts per archive.

Most texts (55) in the corpus stem from the so-called Tomsk school. They are archived at the Tomsk State Pedagogical University, in the **Dul'zon Archives**. Another substantial part of texts (34) originates in the handwritten part of the archive of Angelina I. Kuz'mina (1934–2002). The archive is located at the **Institute of Finno-Ugric / Uralic Studies** at the University of Hamburg.

Five texts originated from Alexander Matthias Castrén (1813–1852); they are archived at the **Department of General Linguistics** in Helsinki.

Three texts were recorded by Kai Donner (1888–1935), two of them are archived at the **Finno-Ugrian Society**, and the archive of the last one is unknown.

Figure 4 *Distribution of texts per archive*



4.7 Researchers

In the corpus, recordings of the Selkup were put together by researchers from different epochs. In the following are some comments on the individual researchers and their records.

The oldest recordings date back to the Finnish philologist and ethnologist **Matthias A. Castrén** (1813–1852). The five heroic songs in the corpus (song) come from his three-year research trip to Siberia (1845–1848).

Nikolaj Grigorovsky's origin is uncertain (Russian or Selkup). What is certain is that he mastered the Selkup language (Hajdú 1973). His "First Selkup reading book" was published in 1879 by the Pravoslav Mission Society in Kazan'. In addition to conventional translations, it also contains four Selkup and three local original stories that he himself collected and provided with Russian translations. The four Selkup stories are represented in the SLC corpus.

The corpus contains two translations (trans) and a folkloric text (flk) by the Finnish linguist and ethnologist **Kai Donner** (1888–1935), which were recorded in the 1910s.

In the 1950s, **Toivo V. Lehtisalo** (1887–1962) recorded some North Selkup texts (flk) on the Turukhan River.

A large part of the texts comes from the so-called '**Tomsk School**'. It was founded by the German-Russian linguist **Andrej P. Dul'zon** (1900–1972). In the 1950s in particular, he researched the various Selkup dialects. His students include a number of Russian linguists who published a large number of Selkup texts in the 1960s – 1980s. Particular mention must be made here of E. G. Bekker, A. W. Bajdak, W. W. Bykonia, N. W. Denning, Sh. Kuper, N. P. Maksimova, Ju. A. Morev, as well as N. A. Tuchkova. The SLC corpus contains 57 texts (flk, nar) recorded by them.

Angelina I. Kuz'mina (1924–2002) was also a student of Dul'zon. Her recordings (text and sound material) covering all Selkup dialects and collected between 1962 and 1972 are archived at the Institute for Finno-Ugric / Uralic Studies at the University of Hamburg. 34 (already published) texts were included in the corpus.

In the 1960s, **László Szabó** recorded 8 Central Selkup texts of the Tym dialect as part of his teaching activity in today's St. Petersburg, all of which were taken over into the SLC corpus.

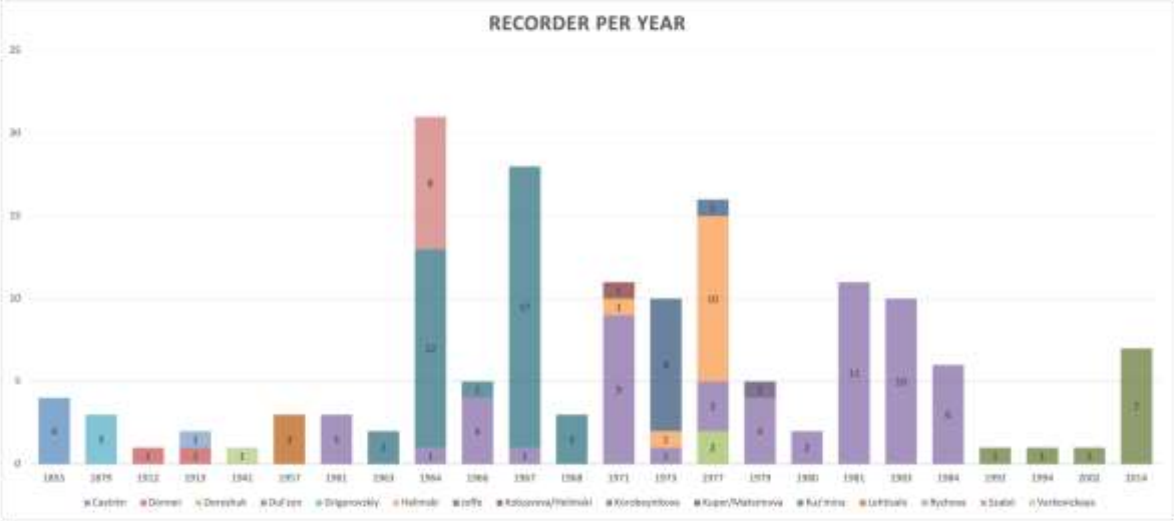
The Russian linguist **Eugen A. Helimski** (1950–2007) was Professor of Finno-Ugric Studies and Linguistics at the University of Hamburg between 1998–2007; 6 texts collected by him in the 1970s were included in the corpus

Texts recorded by the Selkup **Irina A. Korobejnikova** are not original oral ones but translations from different Central and Southern Selkup texts into Korobejnikova's own Selkup dialect (Narym), the corpus contains 5 of these translations.

L. Ju. Joffe, **V. A. Doroshuk** and **E. Ju. Ryzhova** are three Russian researchers who recorded Northern Selkup texts in the 1970s, 12 of them are included in the corpus..

L. Varkovickaya is also a Russian researcher who recorded Northern Selkup texts in the 1940s, 1 text is included in the corpus..

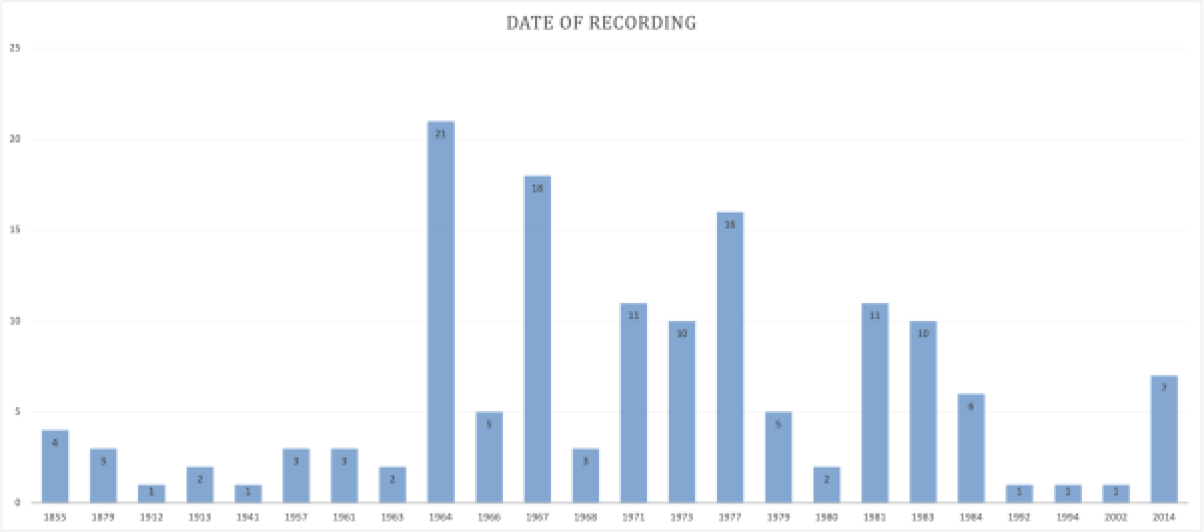
Figure 5 *Distribution of texts per date of recording and recorder*



4.8 Texts

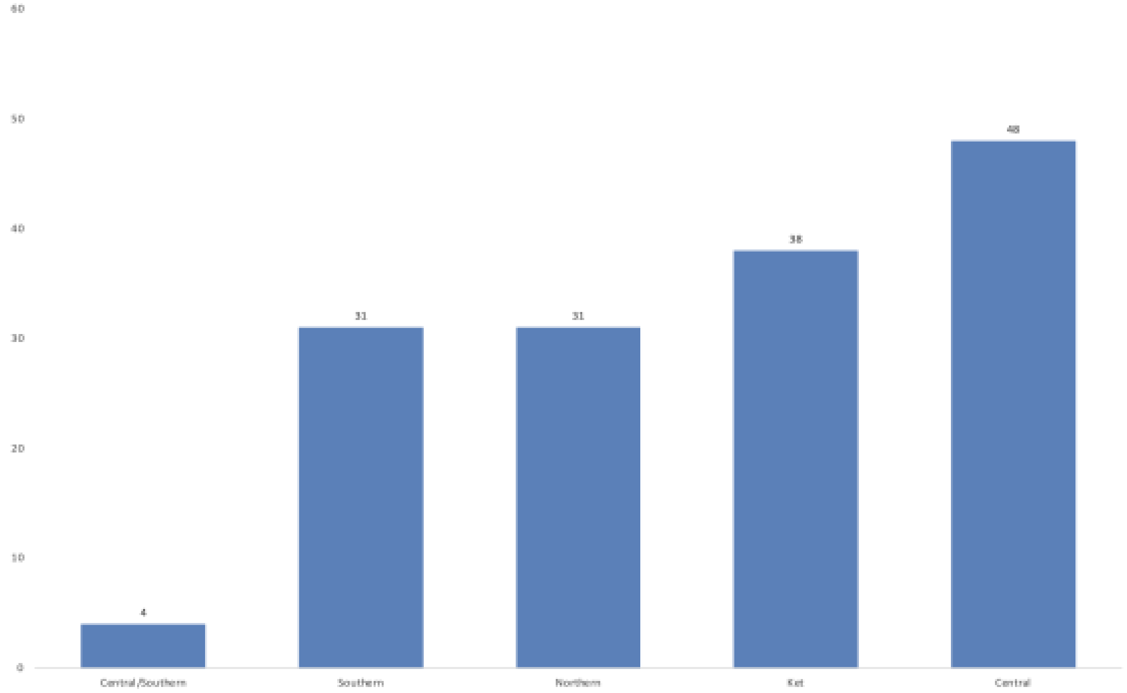
Figure 6 shows the date of recording of the individual texts. It can be seen that texts of the *Selkup Language Corpus* were collected over a period of about 150 years. Most texts date from the 1960s – 1980s, collected by the members of the 'Tomsk school' (see section above). Only some texts in the corpus were collected between 1855 and the beginning of the 1960s or after the 1980s.

Figure 6 *Distribution of texts per date of recording*



When looking for distribution of texts per Selkup dialect group, figure 7 shows that the corpus is not that well-balanced when only looking at the individual texts without taking the length of the texts into account. There are significantly more texts in Southern dialects (69, here split into Ket and Southern) than in Central (48) and Northern (31) ones. Only a few texts – the four heroic songs collected by Castrén – originate from the mixed dialect group.

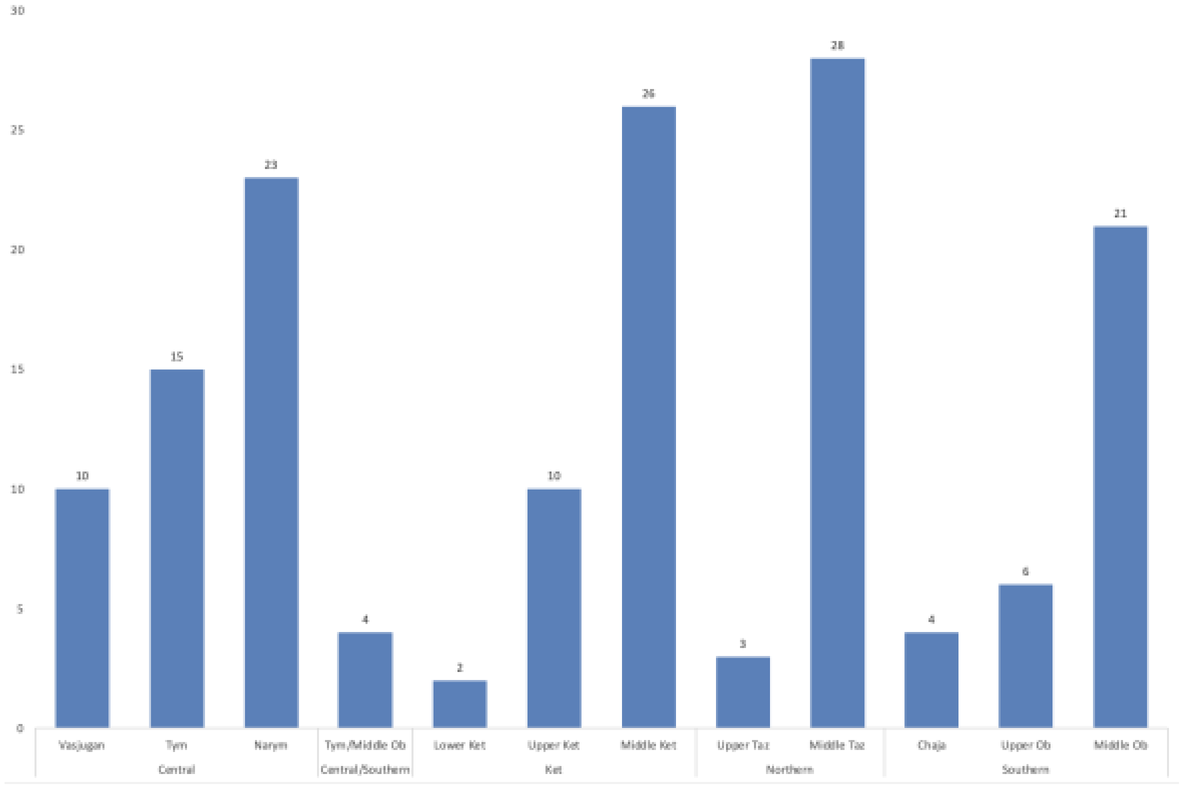
Figure 7 *Distribution of texts per dialect group*



Within the three dialect groups, the distribution is much more different, particularly within the Northern and Southern dialect groups. There are 9 times more texts from the Middle Taz dialect (28) than from Upper Taz (3). Most texts in the Southern dialect group belong to the Middle Ket' dialect

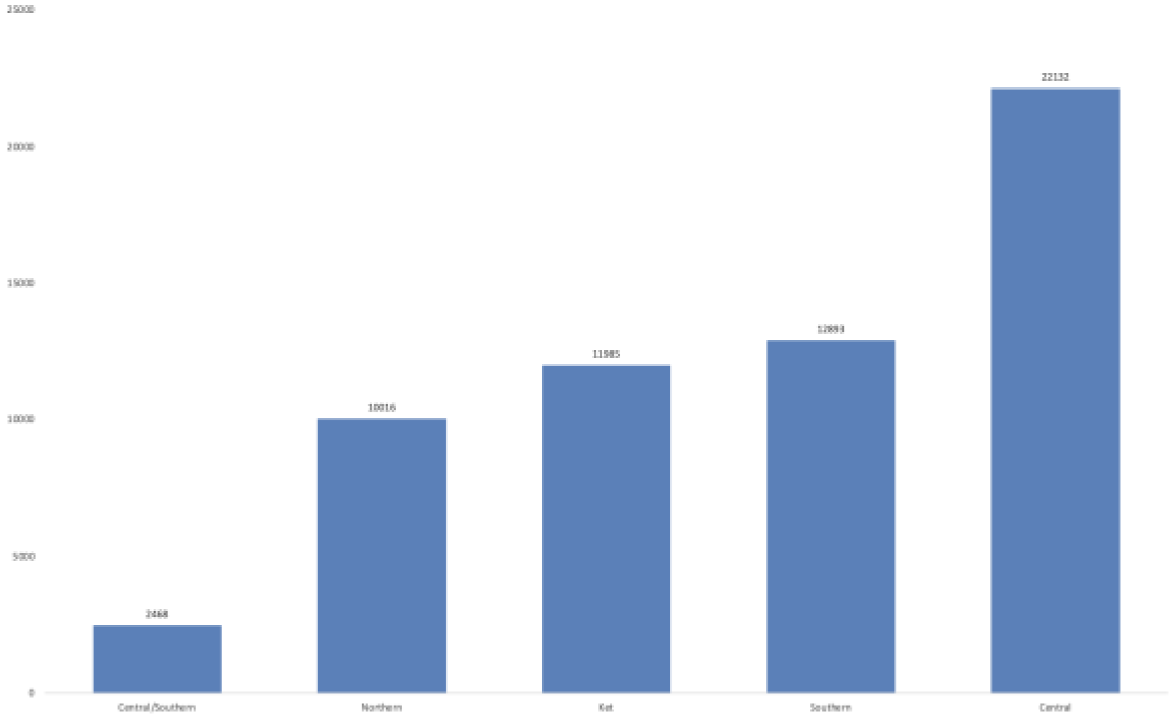
(26). There are significantly less texts in Upper (10) and Lower Ket' (2). The picture is similar in the other Southern dialects: most texts stem from the Middle Ob dialect (21). Much fewer texts are from Upper Ob (6) and Chaya dialect (4). In Central Selkup, most texts are in the Narym dialect.

Figure 8 *Distribution of texts per dialect*



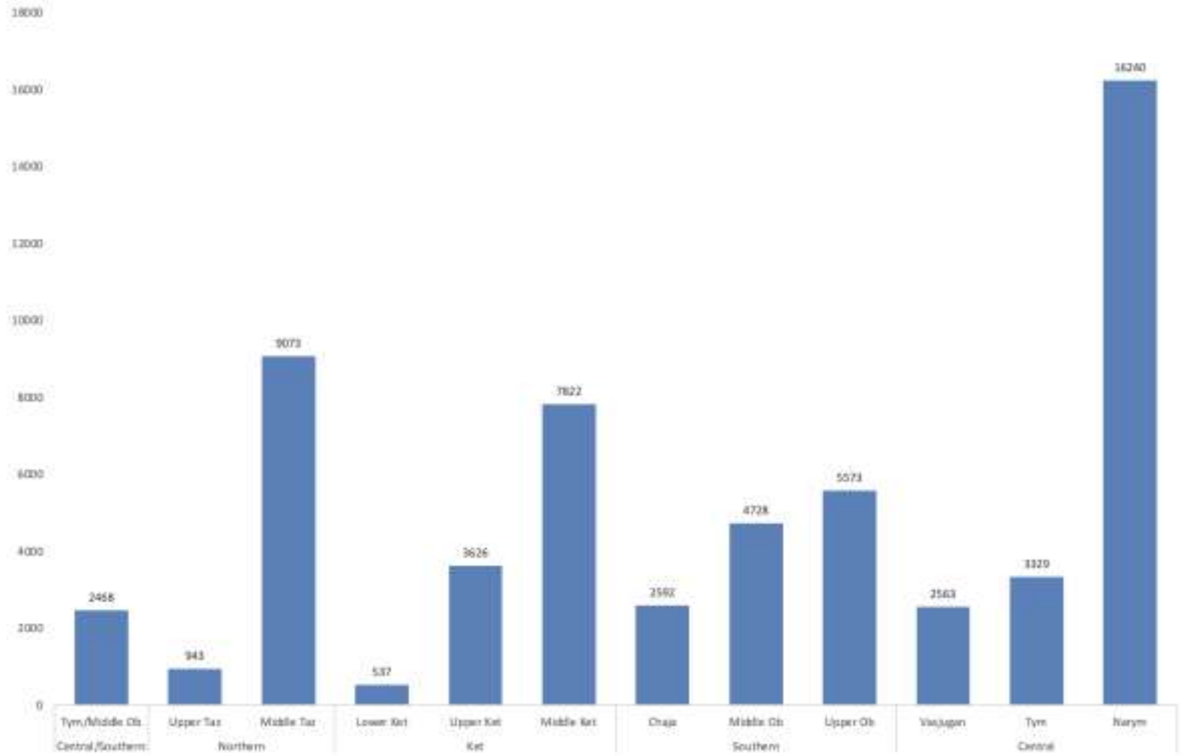
The number of tokens per text varies. Figure 9 shows the distribution of tokens per dialect group. Although there is a big difference in the distribution of texts per dialect, almost the same number of tokens can be found in the Central Selkup (22.132) and Southern Selkup (24.878, here split into Ket' and Southern dialects).

Figure 9 *Distribution of tokens per dialect group*



Within the dialect groups again there is a big difference in distribution of tokens. Figure 10 shows this distribution. Within the Northern dialect group most tokens can be found in the Middle Taz dialect (9.073). Within the Southern Selkup dialect group the distribution is rather well balanced (Chaya: 2.592, Middle Ob: 4.728, Upper Ob: 5.573). Within Ket' and Central dialects the difference is significantly larger: Lower Ket': 537 – Middle Ket': 7.822, Vasyugan: 2.563 – Narym: 16.240 tokens.

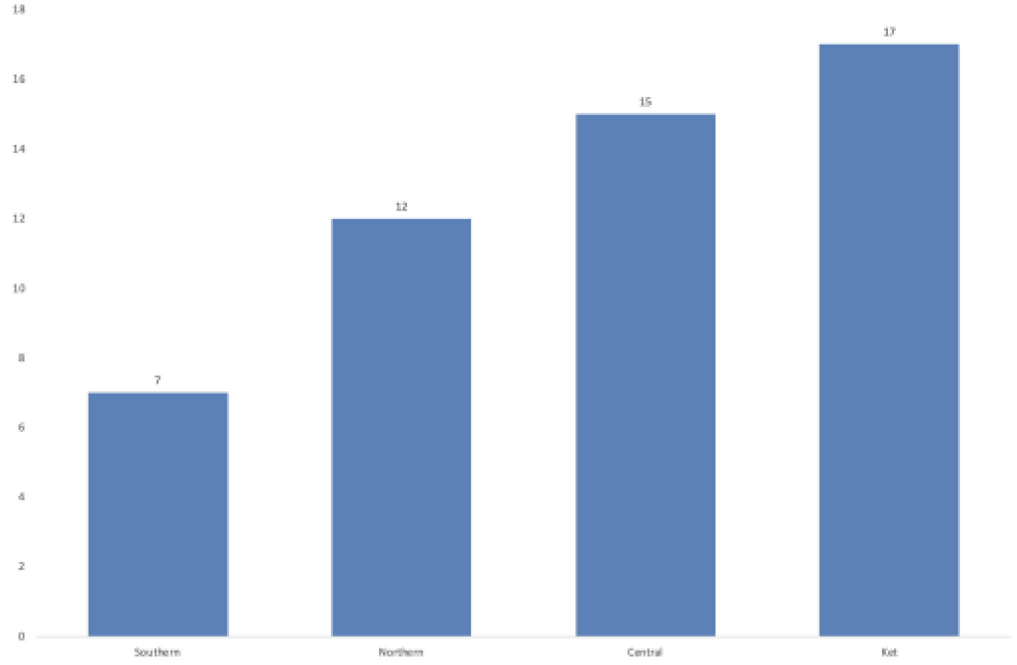
Figure 10 *Distribution of tokens per dialect*



4.9 Speakers

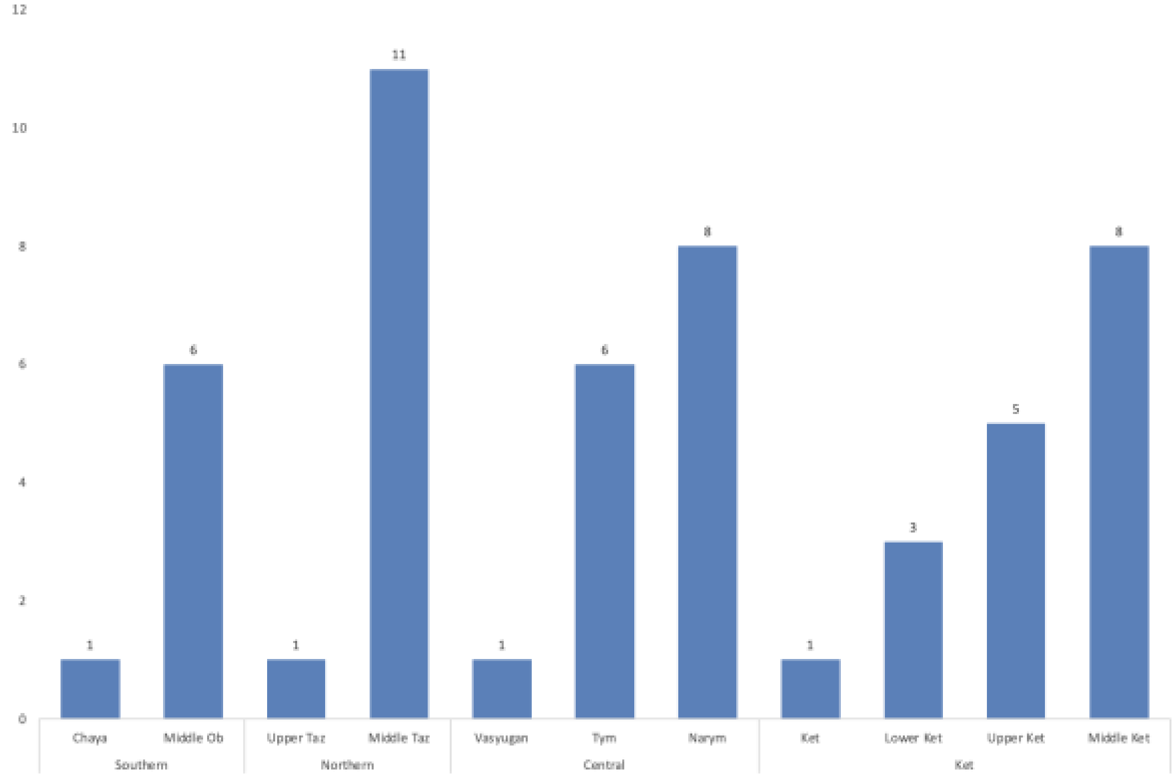
In total, there are 51 Selkup speakers in the *Selkup language corpus*. Figure 11 shows the distribution of speakers per dialect group. The number of speakers is rather well balanced, 17 speakers from Ket' dialects, 15 speakers from Central dialects and 12 speakers from Northern dialects. Only the Southern dialects have fewer speakers in the corpus, as there are only 7 individual speakers.

Figure 11 *Distribution of speakers per dialect group*



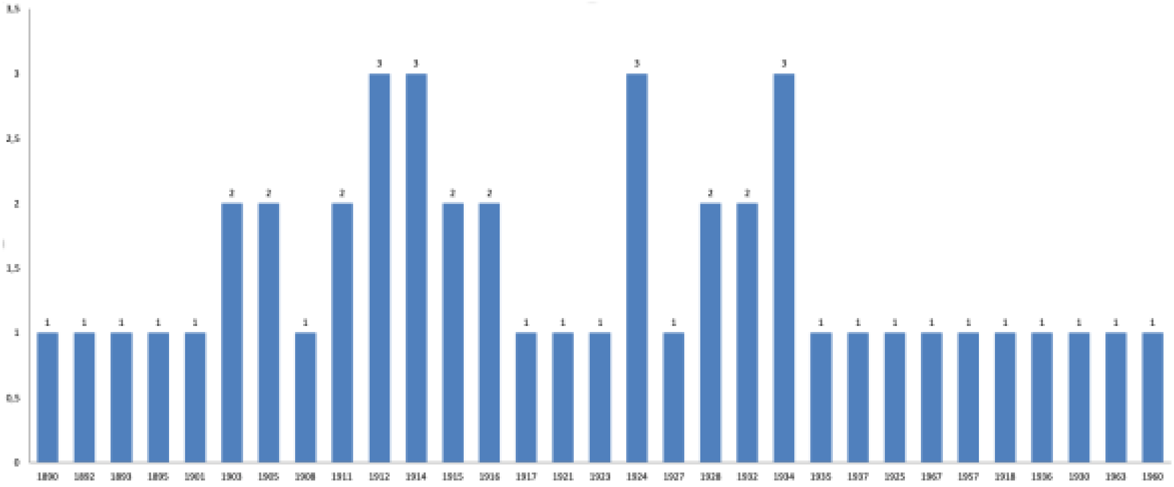
It should be noted that for some dialects, there is only one single speaker, as can be seen in Figure 12 (Chaya, Upper Taz, Vasyugan). All statements about these specific subdialects then refer only to one single speaker. One speaker from the Ket' dialects cannot be precisely attributed.

Figure 12 *Distribution of speakers per dialect*



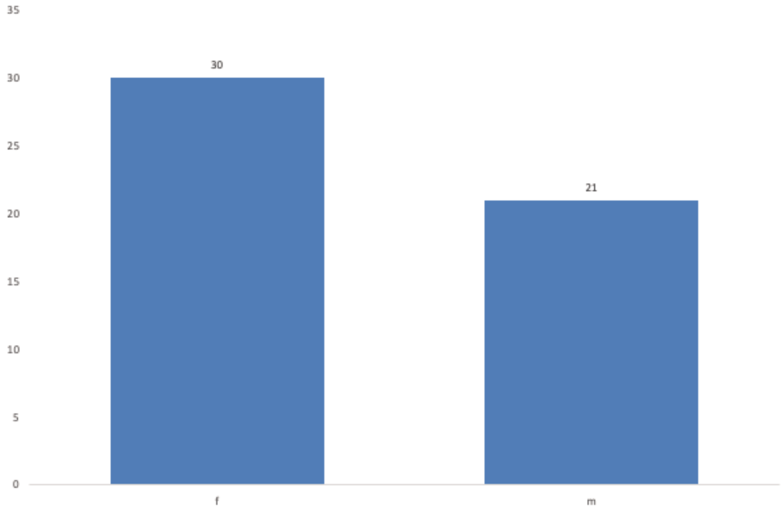
The 51 Selkup speakers in the corpus were born between 1890 and 1960. Figure 13 shows the distribution of speakers per date of birth. The birthday of 4 speakers is unknown, they are not represented in the figure.

Figure 13 *Distribution of speakers per date of birth*



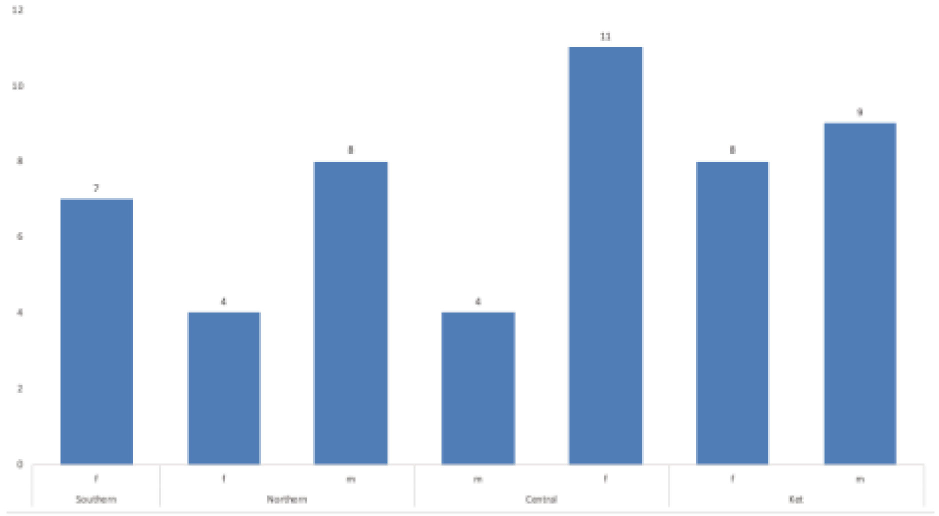
The distribution of speakers per gender is shown in Figure 14. In total there are more male (30) than female speakers (21).

Figure 14 *Distribution of speakers per gender*



Within the dialect groups the picture is different: In the Southern dialects there are no masculine speakers, in the Central dialects there are 3 times more female speakers than male ones. In the Ket' dialects the distribution is well balanced: 8 female to 9 male speakers.

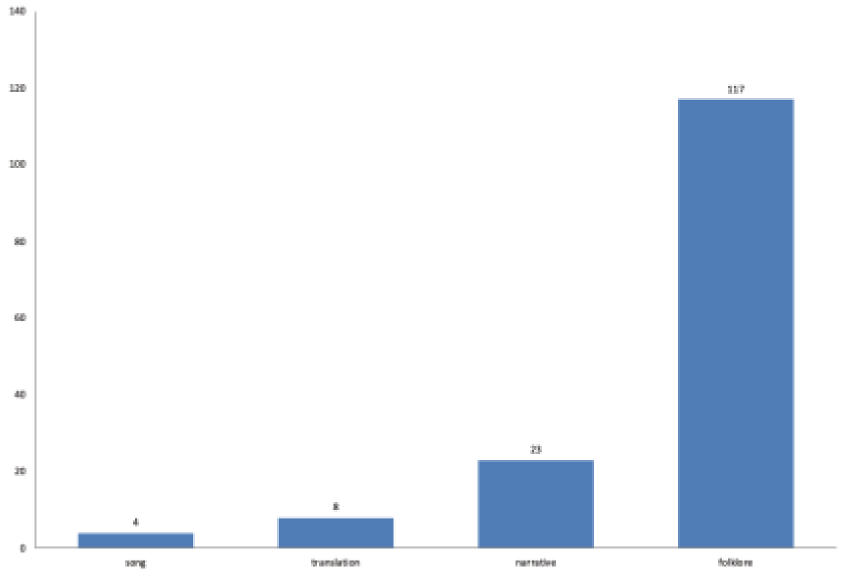
Figure 15 *Distribution of gender per dialect group*



4.10 Genre

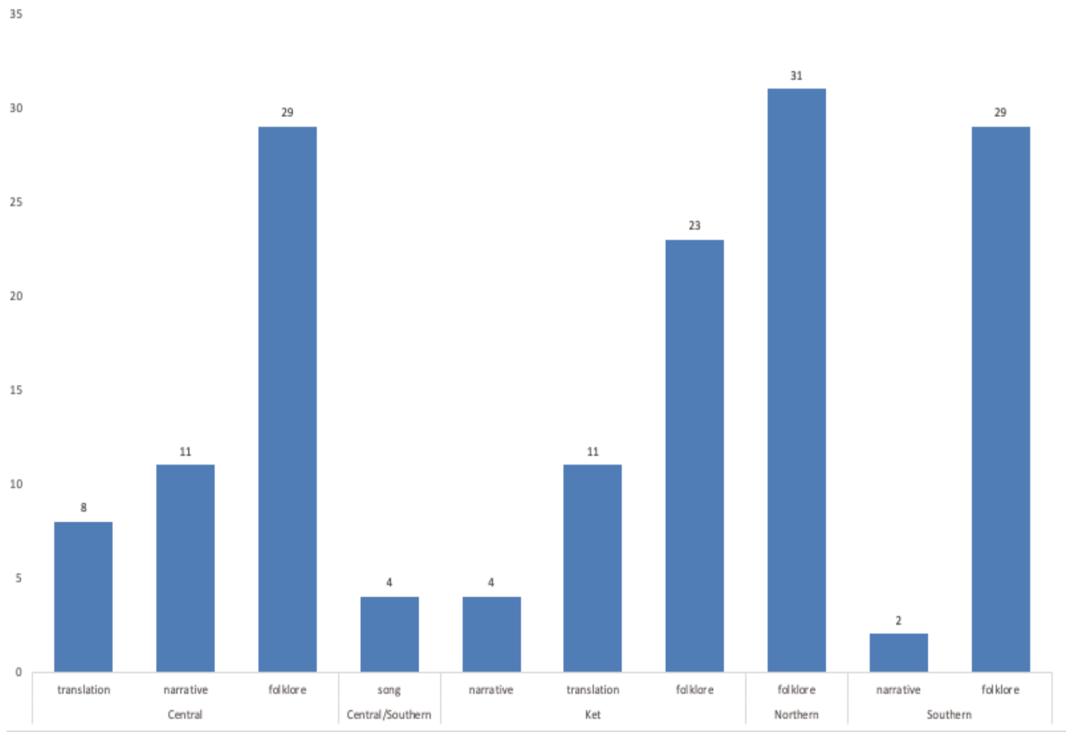
Texts in the Selkup language corpus represent 4 genres: translations (trans), folklore (flk), narrative (nar) and heroic songs (song). Most of the 152 texts in the corpus are folklore ones (117). In contrast, there are 23 narrative texts, 8 translations from Russian into Selkup and 4 heroic songs (all from the mixed dialect group). Figure 16 shows the distribution of texts per genre.

Figure 16 *Distribution of texts per genre*



Within the dialect groups the picture differs as is shown in figure 15: In the mixed dialect group there are only songs, in the Northern dialect group there are only folklore text (31), no narratives nor translations. In Southern dialects (excluding Ket' dialects) there are primarily folklore texts (29), only 2 narrations. Only in the Central and Ket' dialects translations (Central: 8, Ket': 11) as well as narrative (Central: 11, Ket': 4) and folklore texts (Central: 29, Ket': 23) can be found.

Figure 17 *Distribution of texts per genre within dialect group*



5.Structure of the corpus: Transcription and annotation

5.1 Orthography in the corpus

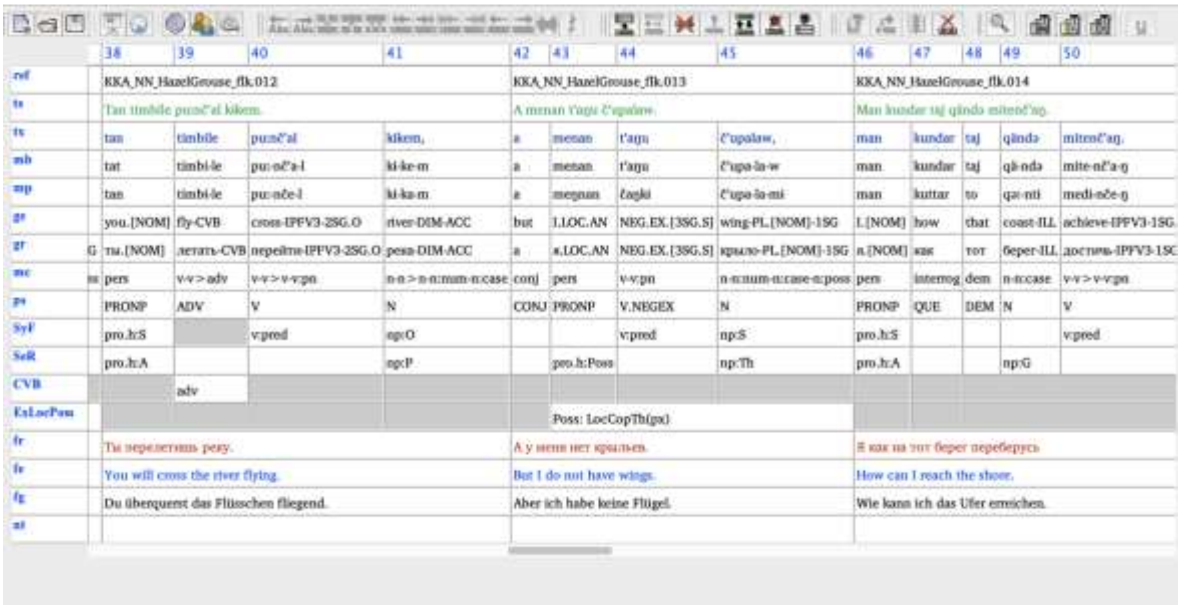
The corpus is a compilation of published texts gathered by several researchers written in either Cyrillic or Latin letters. In the corpus a unified Latin based script is used in the transcription tier (the characters used can be seen in Appendix A). Vowel length is marked with the IPA symbol <ː>, palatalization with <'>. The *charis SIL font* is used throughout the corpus.

5.2 Methodology

All texts are in the EXMARaLDA format, the EXMARaLDA program suite contains various programs: the Partitur editor, EXAKT (search and analysis tool) and CoMA (corpus manager, see section 4.5). The transcripts are all morphologically glossed; the morphological glossing (tier *ge* and *gr*) and the part of speech tagging for each morpheme (tier *mc*) are done in FLEx. The texts are then converted to EXMARaLDA where the remaining annotations are done with the EXMARaLDA Partitur editor.

All texts contain an English translation, most of the texts also contain a Russian and German translation. All language examples contain annotations on the syntactic functions (SyF) and semantic roles (SeR). There are also additional annotations for some texts. Figure 18 shows a typical Selkup text example in the Partitur editor with the different transcription and annotation lines. This also contains annotations on existential, possessive and locative clauses (ExLocPoss) as well as on conversions (CVB). The tiers are described in detail in the following subsections.

Figure 18 *Partitur editor screenshot*

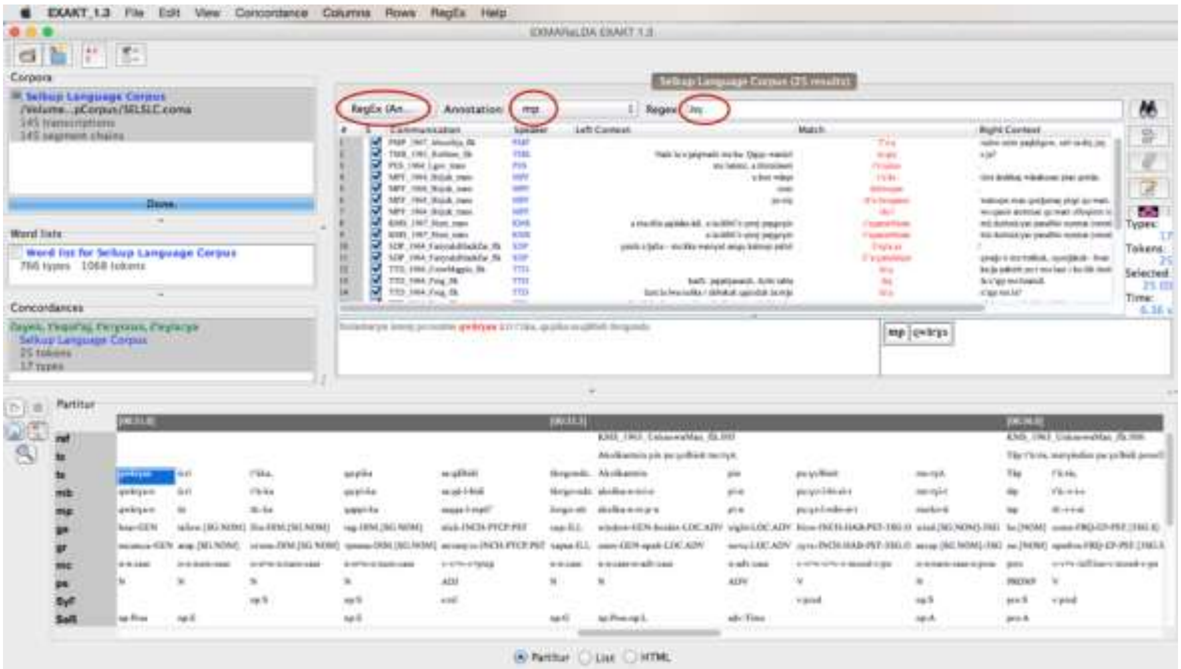


		38	39	40	41	42	43	44	45	46	47	48	49	50
ref		KKA_NN_HazelGrouse_flk.012				KKA_NN_HazelGrouse_flk.013				KKA_NN_HazelGrouse_flk.014				
ts		Tan timbile puːnč'al kikem.				A menan t'agu č'upalaw.				Man kundar taj qində mitenč'aŋ.				
tx		tan	timbile	puːnč'al	kikem,	a	menan	t'agu	č'upalaw,	man	kundar	taj	qində	mitenč'aŋ.
mb		tat	timbi-le	puːnč'a-l	ki-ke-m	a	menan	t'agu	č'upa-la-w	man	kundar	taj	qə-ndə	mite-nč'a-ŋ
mp		tan	timbi-le	puːnče-l	ki-ka-m	a	meŋnan	čaŋki	č'upa-la-mi	man	kuttar	to	qaː-nti	medi-nče-ŋ
ge		you.[NOM]	fly-CVB	cross-IPFV3-2SG.O	river-DIM-ACC	but	I.LOC.AN	NEG.EX.[3SG.S]	wing-PL.[NOM]-1SG	I.[NOM]	how	that	coast-ILL	achieve-IPFV3-1SG.
gr	G	ты.[NOM]	летать-CVB	перейти-IPFV3-2SG.O	река-DIM-ACC	а	я.LOC.AN	NEG.EX.[3SG.S]	крыло-PL.[NOM]-1SG	я.[NOM]	как	тот	берег-ILL	достичь-IPFV3-1SG
mc	m	pers	v-v>adv	v-v>v-v:pn	n-n>n-n:num-n:case	conj	pers	v-v:pn	n-n:num-n:case-n:poss	pers	interrog	dem	n-n:case	v-v>v-v:pn
ps		PRONP	ADV	V	N	CONJ	PRONP	V.NEGEX	N	PRONP	QUE	DEM	N	V
SyF		pro.h:S		v:pred	np:O			v:pred	np:S	pro.h:S				v:pred
SeR		pro.h:A			np:P		pro.h:Poss		np:Th	pro.h:A			np:G	
CVB			adv											
ExLocPoss							Poss: LocCopTh(px)							
fr		Ты перелетишь реку.				А у меня нет крыльев.				Я как на тот берег переберусь				
fe		You will cross the river flying.				But I do not have wings.				How can I reach the shore.				
fg		Du überquerst das Flüsschen fliegend.				Aber ich habe keine Flügel.				Wie kann ich das Ufer erreichen.				
nt														

The corpus compiled with the corpus manager Coma (see also section 4.5) can be analyzed at any time using the EXAKT search and analysis tool (cf. Schmidt & Wörner 2005, Wörner 2010). EXAKT enables the search for (syntactic) phenomena on different levels: The search queries can either refer to transcribed material (transcription search), to descriptions (description search) or to annotations (annotation search). Metadata can also be included in the search. The search results can be evaluated in their respective context.

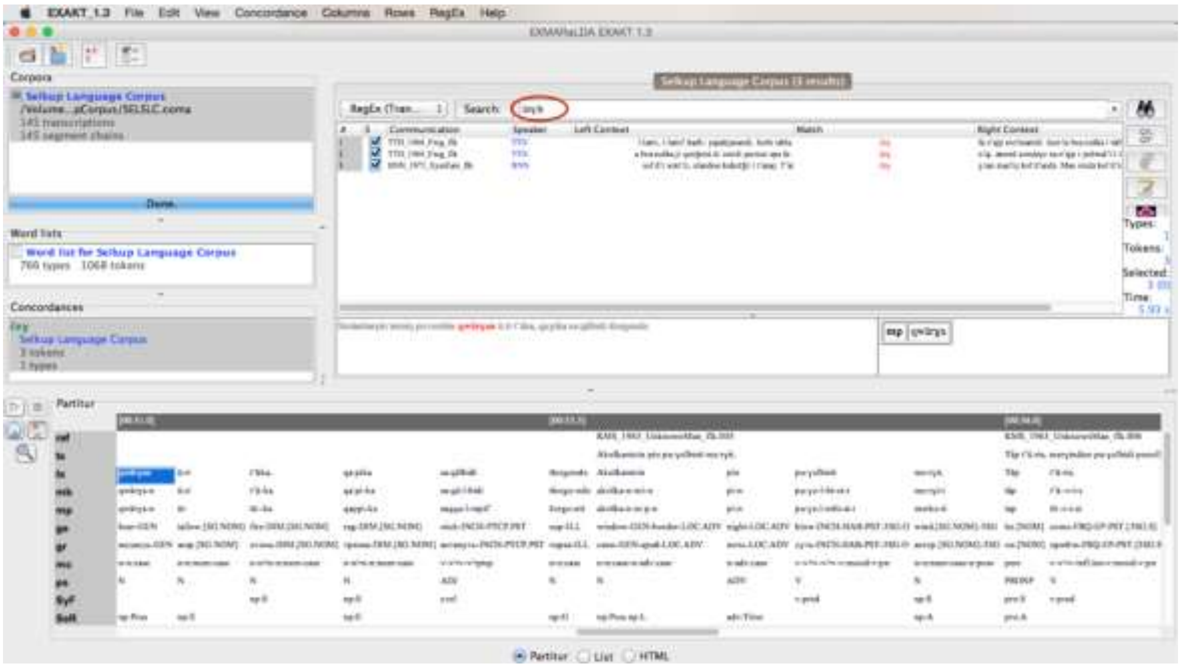
Figure 19 shows the search result of a search expression consisting of simple strings in the form of a KWiC8 concordance. The search term is specified within the context of the reference. If the individual search result is selected, the corresponding transcriptions are displayed in the Partitur editor.

Figure 19 *Exakt screenshot*



Another way to search for a corpus is to search with the help of regular expressions. Here, too, the search results are presented as a KWIC concordance. More complex search queries can also be made with the help of regular expressions. In Figure 20, the regular expression “čeɣ\b” was used to search for all places in the corpus in which the lexical word “čeɣ” occurs.

Figure 20 *Exakt screenshot: search with regular expression*



5.3. Transcription tier and annotation tiers

Each transcription contains at least 12 tiers. The tiers are presented in Table 3.

Table 3 *Tiers in EXMARaLDA Partitur editor*

Name of tier	Description	Type	Category
ref	name of the communication	annotation	obligatory
tx	interlinearization	transcription	obligatory
mb	morpheme break	annotation	obligatory
mp	morphophonemes, underlying form	annotation	obligatory
ge	morphological glossing: English	annotation	obligatory
gr	morphological glossing: Russian	annotation	obligatory
mc	part of speech for each morpheme	annotation	obligatory
ps	part of speech for each word	annotation	obligatory
SyF	syntactic functions	annotation	obligatory
SeR	semantic roles	annotation	obligatory
CVB	converb	annotation	optional
IST	information status	annotation	optional
BOR	borrowing	annotation	optional
ExLocPoss	existential, locative and possessive sentences	annotation	optional
fr	free translation: Russian	annotation	optional
fe	free translation: English	annotation	obligatory
fg	free translation: German	annotation	optional
fh	free translation: Hungarian	annotation	optional
fr_ed	edited free translation: Russian	annotation	optional
fe_ed	edited free translation: English	annotation	optional
fg_ed	edited free translation: German	annotation	optional

nt	notes	annotation	optional

5.3.1 ref – Reference

The tier *ref* gives information about the name of the text and the number of the sentence can also be found here. The tier is of type annotation and obligatory.

5.3.2 ts – Source text

The tier *ts* contains the sentence as it was presented in the source. If there is an audio recording, it is aligned with this tier. The tier is of type annotation, obligatory and always marked in green.

(1)	**ref**	ChDN_1983_HerosDaughter_flk.001
	ts	Ugon ir wargɨmba madet puʒogɨt matur.

5.3.3 tx – Transcription

The tier *tx* is the basis for the morphological glossing, each cell contains one word. The tier is of type transcription, obligatory, linked to the speaker and marked in blue.

(2)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	ts	Ugon ir wargɨmba madet puʒogɨt matur.					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur

5.3.4 mb – Morpheme breaks

The tier *mb* shows a morpheme by morpheme break-up of the words, the morphemes are separated by hyphens; zero morphemes are left out in this tier. The tier is of type annotation and obligatory.

(3)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur

5.3.5 mp – Morphophonemes

In the tier *mp* the underlying form of all morphs is presented. Selkup is a language with complex morphophonological processes; hence words can have many allomorphs. Furthermore, Selkup is a non-standardised language with a vast dialectical continuum. Morphs may occur in several written forms. The tier is of type annotation and obligatory.

(4)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur
	mp	ugon	iːr	wargɨ-mbɨ	maǯ’o-n	puǯo-qɨn	matur

5.3.6 gr, ge – Russian and English morpheme glossing

The tiers *gr* and *ge* are for the interlinear morpheme-by-morpheme glossing. The lexical meaning of the stem is given in either Russian or English, the glossing labels are the same for both languages, the Latin script is used here. The labelling follows international standards (mostly the Leipzig Glossing Rules); the additions made to this basic label set can be found in Appendix 1.

A dot shows that two (or more) components semantically belong together and is also used to separate stems in compounds, a dash separates alternative meanings, and square brackets indicate non-overt morphemes. Combinations of person and number markings are combined in one gloss without a dot: e.g. 1PL for first person plural.

The unmarked category simple singular is only marked if the word is in nominative, then a complex gloss is used: [SG.NOM], apart from that singular is not marked in the corpus.

Selkup has two types of conjugation: a subjective and an objective, in the glossing, this is marked by .S or .O following the person of the verb, in the plural the forms collapsed and are hence marked by S/O. The tiers are of type annotation and obligatory.

(5)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur
	mp	ugon	iːr	wargɨ-mbɨ	maǯ’o-n	puǯo-qɨn	matur
	ge	earlier	long.ago	live-PST.REP.[3SG.S]	taiga-GEN	inside-LOC	hero.[SG.NOM]
	gr	раньше	давно	жить-PST.REP.[3SG.S]	тайга-GEN	внутренность-LOC	герой.[SG.NOM]

5.3.7 mc – Morpheme class

The tier *mc* is used to indicate the morphological category of each morph – the part of speech of the lexical stem (see Table 4) and the category of the suffixes (see Table 5).

Table 4 *Tags of lexical stems*

tag	description	tag	description
adj	adjective	ptcp	participle
adv	adverb	ptcl	particle
clit	clitic	pers	personal pronouns
conj	conjugation	pp	postposition
dem	demonstrative	pro	pronoun
emph	emphatic pronouns	quant	quantifier
interj	interjection	v	verb
interrog	interrogative pro-form		

n	noun		
num	numeral		

Table 5 *Tags for inflection*

category	tag	description
nominal	num	number
	case	case
	poss	possessor
verbal	tense	tense
	mood	mood
	pn	personal ending

(6)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur
	mp	ugon	iːr	wargɨ-mbɨ	maǯ'o-n	puǯo-qɨn	matur
	ge	earlier	long.ago	live-PST.REP.[3SG.S]	taiga-GEN	inside-LOC	hero.[SG.NOM]
	gr	раньше	давно	жить-PST.REP.[3SG.S]	тайга-GEN	внутренность-LOC	герой.[SG.NOM]
	mc	adv	adv	v-v:mood-v:pn	n-n:case	n-n:case	n-n:case

5.3.8 ps – Part of speech

In the tier *ps* part of speech for each word form is tagged. The categorization is syntax oriented. Some classes are divided into subcategories: nouns are divided into common and proper nouns, particles, auxiliaries and verbs are divided into affirmative and negative categories with a special tag for the negative existential verb.

Cardinal numbers belong to the category QUANT while ordinal numerals are annotated as adjectives.

The pronominal class is split up: interrogative pronominals are tagged as QUE, adverbial pronominals as ADV, demonstrative pronouns as DEM and personal pronouns as PRONP. Personal pronouns in the function of a possessive pronoun are tagged with PRONPOS, pronouns being neither personal nor possessive are not further split and only marked as pronouns.

Table 6 *Tags for part of speech*

tag	description	tag	description
N	common noun	INDF	indefinite
NPR	proper noun	INTS	intensifier
NUM	numeral	INTERJ	interjection
PRON	pronoun	NPI	negative polarity item
PRONP	personal pronoun	ONOM	onomatopoeia
PRONPOS	possessive pronoun	PTCL	affirmative particle
ADJ	adjective	PTCL.NEG	negative particle
ADV	adverb	PREP	preposition
V	affirmative verb	PP	postposition
V.NEGEX	negative existential verb	PREV	preverb
CONJ	conjunction	QUANT	quantifier
DEM	demonstrative	QUE	question word

(7)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur
	mp	ugon	iːr	wargɨ-mbɨ	maǯ’o-n	puǯo-qɨn	matur
	ge	earlier	long.ago	live-PST.REP.[3SG.S]	taiga-GEN	inside-LOC	hero.[SG.NOM]
	gr	раньше	давно	жить-PST.REP.[3SG.S]	тайга-GEN	внутренность-LOC	герой.[SG.NOM]
	mc	adv	adv	v-v:mood-v:pn	n-n:case	n-n:case	n-n:case
	ps	ADV	ADV	V	N	N	N

5.3.9 fr, fe, fg, fh – Free Translations into Russian, English, German and Hungarian

The tiers *fr*, *fe*, *fg,* and *fh* are used for free translations into Russian, English, German and Hungarian. The English translation (*fe*) is obligatory for all texts, a Russian translation (*fr*) is provided for most texts (it is marked in red); a German translation (*fg*) as well as Hungarian (*fh*) is added if available.

(8)	ref	ChDN_1983_HerosDaughter_flk.001					
	ts	Ugon ir wargɨmba madet puʒogɨt matur.					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur
	mp	ugon	iːr	wargɨ-mbɨ	maǯ’o-n	puǯo-qɨn	matur
	ge	earlier	long.ago	live-PST.REP.[3SG.S]	taiga-GEN	inside-LOC	hero.[SG.NOM]
	gr	раньше	давно	жить-PST.REP.[3SG.S]	тайга-GEN	внутренность-LOC	герой.[SG.NOM]
	mc	adv	adv	v-v:mood-v:pn	n-n:case	n-n:case	n-n:case
	ps	ADV	ADV	V	N	N	N
	fr	Давным - давно жил в чаще леса богатырь.					
	fe	Long ago, a hero lived in the forest.					
	fg	Vor langer Zeit lebte ein Held im Wald.					

5.4 Annotation of Syntactic Function (SyF)

In the tier *SyF* the core syntactic functions subject, object and predicate are annotated. The tier is of type annotation and is obligatory. The annotation scheme corresponds with the scheme used for annotation of Nganasan corpus (Wagner-Nagy et al. 2018). The form of annotation is <form.function>

Table 7 *Tags for core syntactic functions*

tag	description
S	subject
O	object
pred	predicate

5.4.1. Annotation of subject

The subject usually is in nominative, and either a common noun, a proper noun or a pronoun. But also adjectives can function as subjects. If the subject is human or an anthropomorphised animal, it is marked human with <h>. As Selkup is a pro-drop language, the subject can be marked solely on the verb; it is therefore useful to annotate also covert subjects. In that case, the dropped referent and predicate are annotated in the same cell.

Table 8 *Tags for subjects*

tag	description
np:S	nominal subject

pro:S	pronominal subject
np.h:S	nominal human subject
pro.h:S	pronominal human subject
0.1:S	dropped first person subject
0.2:S	dropped second person subject
0.3:S	dropped third person subject
0.1.h:S	dropped human first person subject
0.2.h:S	dropped human second person subject
0.3.h:S	dropped human third person subject

A dropped referent is shown in example (9)

(9)	**ref**	ChDN_1983_MistressOfFire_flk.089			
	ts	Hel’d’ po:p pellag’e šogort.			
	tx	Hel’d’	po:p	pellag’e	šogort.
	mb	hel’d’	po:-p	pel-la-g’e	šogor-t
	mp	hel’d’	po-m	pat-lä-k	šoɣor-ntɨ
	ge	seven	tree-ACC	put-OPT-1SG.S	stove-ILL
	gr	семь	дерево-ACC	положить-OPT-1SG.S	печь-ILL
	mc	num	n-n:case	v-v:mood-v:pn	n-n:case
	ps	QUANT	N	V	N
	SyF		np:O	0.1.h:S v:pred	
	fr	Семь поленьев положу в печь.			
	fe	I put seven logs in the stove.			

5.4.2. Annotation of object

Direct objects in Selkup are usually marked with accusative but can be found in e.g. nominative as well. Human objects (or humanlike animals) are marked as human with <h>.

Table 9 Tags for objects

tag	description
np:O	nominal object
pro:O	pronominal object
np.h:O	nominal human object
pro.h:O	pronominal human object
0.3:O	dropped third person object
0.3.h:O	dropped human third person object

Example (10) shows a direct object in accusative case.

(10)	**ref**	ChDN_1983_HerosDaughter_flk.008				
	ts	Tab wargɨ hurum naj kwatkumbad.				
	tx	Tab	wargɨ	hurum	naj	kwatkumbad
	mb	tab	wargɨ	huru-m	naj	kwat-ku-mba-d
	mp	tab	wargɨ	hurup-m	naj	kwat-ku-mbɨ-tɨ
	ge	he.[NOM]	big	wild.animal-ACC	also	kill-ITER-PST.REP-3SG.O
	gr	он.[NOM]	большой	зверь-ACC	тоже	убыть-ITER-PST.REP-3SG.O
	mc	pers-n:case	adj	n-n:case	ptcl	v-v > v-v:mood-v:pn
	ps	PRONP	ADJ	N	PTCL	V
	SyF	pro.h:S		np:O		v:pred
	fr	Она и на крупных зверей охотилась.				
	fe	She also hunted big wild animals.				
	fg	Sie jagte auch große wilde Tiere.				

5.4.3. Annotation of predicate

In the predicate position verbs, nouns, adjectives, participles and converbs can occur; nouns, adjectives and participles can be accompanied by copula but it is not necessarily the case. Converbs have to be accompanied either by copula or an auxiliary.

Table 10 *Types of predicates*

tag	description
v:pred	verbal predicate
n:pred	nominal predicate
adj:pred	attributive predicate
ptcl:pred	particle predicate
cvb:pred	converbal predicate

An example for a nominal predicate, accompanied by copula, is shown by the first part of the sentences in (11):

(11)	ref	ChDN_1983_ItjaStayedAlone_flk.002			
	ts	Ad’ade eppɨmba menertɨl qup […]			
	tx	ad’ade	eppɨmba	menertɨl	qup
	mb	ad’a-de	e-ppɨ-mba	mene-r-tɨl	qup
	mp	aǯ’a-tɨ	eː-mbɨ-mbɨ	mene-r-ntɨl’	qum
	ge	father.[SG.NOM]-3SG	be-HAB-PST.REP.[3SG.S]	hunt-FRQ-PTCP.PRS	human.being.[SG.NOM]
	gr	отец.[SG.NOM]-3SG	быть-HAB-PST.REP.[3SG.S]	охотиться-FRQ-PTCP.PRS	человек.[SG.NOM]
	mc	n-n:case-n:poss	v-v>v-v:mood-v:pn	v-v>v-v>ptcp	n-n:case
	ps	N	V	ADJ	N
	SyF	np.h:S	cop		n:pred
	fr	Отец был охотником, в лес ушел, из леса не вернулся.			
	fe	The father was a hunter, went to the forest, did not return from the forest.			
	fg	Der Vater war Jäger, ging in den Wald, kehrte aus dem Wald nicht zurück.			

5.4.4. Annotation of subordinate clauses

In the annotation of subordinate clauses five types are distinguished: adverbial, conditional, purpose, relative, temporal and complement. Table 11 shows the tagging set for these:

Table 11 *Types of subordinate clauses*

tag	description
s:adv	adverbial
s:cond	conditional
s:purp	purpose
s:rel	relative
s:temp	temporal
s:compl	complement

In example (12) a purpose clause is shown:

(12)	**ref**	ChDN_1983_GirlAndIce_flk.002			
	ts	Podɨp aramum megu t’umba.			
	tx	Podɨp	aramum	megu	t’umba.
	mb	pod-ɨ-p	aramu-m	me-gu	t’u-mba
	mp	p’ed’-ɨ-m	aramu-m	me-gu	tö:-mbɨ
	ge	axe-EP-ACC	icehole-ACC	do-INF	come-PST.REP.[3SG.S]
	gr	топор-EP-ACC	пробурь-ACC	делать-INF	прийти-PST.REP.[3SG.S]
	mc	n-infl:ins-n:case	n-n:case	v-v:ninf	v-v:mood-v:pn
	ps	N	N	V	V
	SyF	np:O	s:purp		0.3.h:S v:pred
	fr	Прорубь сделать пришел.			
	fe	He brought an axe to make an ice hole.			

5.5. Annotation of Semantic Roles (SeR)

The tier SeR is for the annotation of semantic roles. The tier is of type annotation and obligatory. The annotation scheme corresponds with the scheme used for annotation of Nganasan corpus (Brykina et al. 2016). The entry is built using the GRAID principle (Haig & Schnell 2017): <form.animacy:function> with some modifications. For now the following functions are implemented in the corpus.

Table 12 *Tags for semantic roles – functions*

abbreviation	description	comment
A	agent	Initiator of the action, in control of its action – it is causing and responsible for the happening.
B	beneficiary	Entity for whose benefit the action was performed or who is the beneficiary of the state, action or procedure.
Cau	cause	Entity causing the happening.
Com	comitative	Entity that convoys the participant of the action
E	experiencer	Entity that experiences or feels an action and is not responsible or in control of it - emotion, volition, cognition, perception (verbs like: *live, die, see, love, hate, understand, hear, taste, frighten, wish, want, think, remember, feel*)
G	goal	Location or entity towards something is moving
Ins	instrument	Entity by which the action is performed
L	location	Location in which something is situated
P	patient	Undergoer of the action, is changed by the action.
Path	path	Direction something is moving along
Poss	possessor	Entity who possesses something
R	recipient	Entity who receives something Addressee of a verb of speech (verbs like: *give, buy, bring, carry and say, be mad, shout at someone*)
So	source	Place of origin or original owner of something in a transfer
Th	theme	Entity which is moved by some action Entity whose location is specified (e.g. in existential and locative sentences) Entity about which a cognitive, communicative or emotional situation is about.
Time	time	Particular time Interval of time

5.5.1. Form of referent

The form of the referent is marked in the corpus with the following tag set; Selkup is a pro-drop language hence it is advisable to mark also covert referents.

Table 13 *Tags for semantic roles – form of referent*

abbreviation	description
0	covert
adv	adverb
np	nominal phrase
pp	postposition
pro	pronoun

5.5.2. Properties of referent

In the corpus, all three persons are annotated. Furthermore, it is tagged if the referent is human or non-human: a human referent is marked with <h> while a non-human referent is not marked. Anthropomorphized animals are also annotated as human, also groups in which at least one participant is human are marked as human.

Table 14 *Tags for semantic roles – properties*

abbreviation	description	abbreviation	description
1	first person	3	third person
2	second person	h	human referent

(14)	**ref**	ChDN_1983_HerosDaughter_flk.001					
	ts	Ugon ir wargɨmba madet puʒogɨt matur.					
	tx	Ugon	ir	wargɨmba	madet	puʒogɨt	matur
	mb	ugon	ir	wargɨ-mba	made-t	puʒo-gɨt	matur
	mp	ugon	i:r	wargɨ-mbɨ	maǯ'o-n	puǯo-qɨn	matur
	ge	earlier	long.ago	live-PST.REP.[3SG.S]	taiga-GEN	inside-LOC	hero.[SG.NOM]
	gr	раньше	давно	жить-PST.REP.[3SG.S]	тайга-GEN	внутренность-LOC	герой.[SG.NOM]
	mc	adv	adv	v-v:tense-v:pn	n-n:case	n-n:case	n-n:case
	ps	ADV	ADV	V	N	N	N
	SeR	adv:Time	adv:Time		np:Poss	np:L	np.h:E

	SyF			v:pred			np.h:S
	fr	Давным - давно жил в чаще леса богатырь.					
	fe	Long ago, a hero lived in the forest.					
	fg	Vor langer Zeit lebte ein Held im Wald.					

5.6. Annotation of Information Status (IST)

The here used annotation of the information status is a combination of the guidelines taken from Götze et al. (2007) and some elements of the RefLex Scheme (Riester & Baumann 2017, first publications about that in 2014). The scheme was elaborated by Sándor Szeverényi.[9] The three core categories given, accessible and new are kept and further subdivided:

1. A new referent has not been mentioned in the discourse and is completely new to the hearer and cannot be determined via context.
2. A referent is given if mentioned in the discourse beforehand, it is marked as active if mentioned in the clause before; else it is marked as inactive.

The accessibility of a referent is further distinguished in four subcategories: situative (it is clear from the situation that the referent is needed and is therefore accessible), inferable (e.g. my hand, the door of a house), aggregational (two already mentioned referents emerge as one, e.g. my mother, my father – my parents), general (knowledge about the world, e.g. *sun*)

Table 15 *Tags for Information Status*

tag	**description**	**category**
giv	given (underspecified)	given
giv-active	given active	
giv-inactive	given inactive	
accs-sit	accessible situative	accessible
accs-inf	accessible inferable	
accs-agg	accessible aggregational	
accs-gen	accessible general	
new	new	new

Also in that annotation line, zero referents are marked by a leading 0.,e.g. 0.accs-inf, and referents appearing in a direct quotation are marked by a following -Q, e.g. accs-inf-Q. These two markers can also be combined as in e.g. 0.accs-inf-Q.

An example sentence with tagged information status can be seen here:

[9] The first version is described in Brykina et al. 2016. The here used version is published also in Wagner-Nagy et al. 2018.

(15)	**ref**	TTD_1964_Squirrel_nar.004			
	ts	Onek tabet qo:bɨmdɛ n’iŋgle:be.			
	tx	Onek	tabet	qo:bɨmdɛ	n’iŋgle:be.
	mb	onek	tabe-t	qo:bɨ-m-dɛ	n’iŋ-le:-be
	mp	onek	tapäk-n	kobɨ-m-tɨ	n’iŋ-la-m
	ge	myself.[NOM]	squirrel-GEN	skin-ACC-3SG	take.off-FUT-1SG.O
	gr	я.сам.[NOM]	белок-GEN	шкура-ACC-3SG	снять-FUT-1SG.O
	mc	emph	n-n:case	n-n:case-n:poss	v-v:tense-v:pn
	ps	INTS	N	N	V
	SyF	pro.h:S		np:O	v:pred
	SeR	pro.h:A	np:Poss	np:P	
	IST	giv-active	giv-inactive	accs-inf	
	fe	I’ll skin the squirrel myself.			

5.7. Annotation of Borrowing (BOR)

Borrowing is annotated in several tiers: BOR, BOR-Phon and BOR-Morph, the here used schema is taken from the Nganasan Spoken Language Corpus (cf. Brykina et al 2018 and Wagner-Nagy et al. 2018). In the tier BOR the source language and the lexical type is annotated:
RUS: for Russian
DOL: for Dolgan, etc.

Different types of loanwords are annotated according to Myers-Scotton (2002, 2006). It is distinguished between cultural borrowings and core borrowings. A further type is grammatical borrowing such as conjunctions (e.g *i* ‘and’). Additionally, borrowed discourse markers and modal words are annotated. Table 16 below shows the annotation tags for the tier BOR.

Table 16 *Tags for BOR*

Annotation Tag	Description
cult	cultural borrowing
core	core borrowing
gram	grammatical borrowing
mod	modal word borrowed
disc	discourse marker borrowed

During the annotation, structural integration (phonetic/phonological and inflectional) of nouns and verbs is taken into consideration. This phenomenon is annotated in tier BOR-Phon.

Table 17 *Tags for phonological adaptation strategies (Tier BOR-Phon)*

Types of adaptation	**Tag**	**Comment**
deletion	inCdel	initial consonant deletion
	inVdel	initial vowel deletion (aphaeresis)
	medCsdel	medial consonant deletion
	medVdel	medial vowel deletion (syncope)
	finCdel	final consonant deletion
	finVdel	final vowel deletion (apocope)
insertion	inVins	initial vowel insertion
	medVins	medial vowel insertion
	finVins	final vowel insertion
substitution	Csub	consonant substitution
	Vsub	vowel substitution
lenition	lenition	weakening
fortition	fortition	strengthening

In case of verbal borrowings, tier BOR-Morph is used for further annotation, by applying Wohlgemuth's typology (2009). Wohlgemuth differentiates between the following categories:
a) direct insertion (no morphological adaptation)
b) indirect insertion (adaptation by affixation, etc.)
Table 18 shows the annotation tags for the tier BOR-Morph.

Table 18 *Tags for morphological adaptation strategies (tier BOR-Morph)*

Type	**Tag for strategy**	**Tag for inflection**	**Comment**
direct insertion	dir:	bare	direct insertion without any morphological adaptation
	dir:	infl	direct insertion with further inflection
indirect insertion	indir:	bare	insertion with morphological adaptation without further inflection
	indir:	infl	insertion with morphological adaptation with further inflection
paradigm insertion	parad:	bare	the verb is borrowed with verbal inflexion from the donor language, but is not further inflected

	parad:	infl	the verb is borrowed with verbal inflexion from the donor language and is not further inflected

Example (16) shows a Russian borrowing:

(16)	**ref**	ChDN_1983_BearCameIntoVillage_nar.004			
	ts	Man akoškautɛ ponɛ manʒɛdegak.			
	tx	Man	akoškautɛ	ponɛ	manʒɛdegak.
	mb	man	akoška-utɛ	ponɛ	manʒɛ-de-ga-k
	mp	man	akoška-un	po:ne	mantɨ-ntɨ-ŋɨ-k
	ge	I.[NOM]	window-PROL	outward(s)	look-IPFV2-AOR-1SG.S
	gr	я.[NOM]	окно-PROL	наружу	смотреть-IPFV2-AOR-1SG.S
	mc	pers-n:case	n-n:case	adv	v-v>v-v:tense-v:pn
	ps	PRONP	N	ADV	V
	SyF	pro.h:S			v:pred
	SeR	pro.h:E	np:Path		
	BOR		RUS:cult		
	fr	Я из окна на улицу выглянула.			
	fe	I looked out of the window.			

5.8. Annotation of existential, locative and possessive sentences (ExLocPoss)

In the tier *ExLocPoss* existential, locative and possessive sentences and the order of their components are annotated to make the sentences searchable through their word order. The scheme was elaborated by Josefina Budzisch. At first the type of sentences (see Table 19) is indicated after that the order of the components (see Table 20) is marked; possessive suffixes on the theme are only marked in possessive sentences.

Table 19 *Type of sentences*

tag	**description**
Ex	existential sentence
Loc	locative sentence
Poss	possessive sentence

Table 20 *Components of the sentences*

tag	**description**
Th	theme
Loc	location
Cop	copula

(px)	possessive suffix

(17)	**ref**	MNS_1984_BrotherSister_flk.017	
	ts	Hör čaːŋgwa.	
	tx	hör	čaːŋgwa
	mb	hör	čaːŋg-wa
	mp	hɛr	čaŋkɨ-ŋɨ
	ge	snow.[SG.NOM]	NEG.EX-AOR.[3SG.S]
	gr	снег.[SG.NOM]	NEG.EX-AOR.[3SG.S]
	mc	n-n:num-case	v-v:tense-v:pn
	ps	N	V.NEGEX
	SyF	np:S	v:pred
	SeR	np:Th	
	ExLocPoss	Ex: ThCop	
	fe	There is no snow.	

5.9. Annotation of converbal constructions (CVB)

To distinguish the different functions of converbal constructions, converbs and their accompanying finite verbs are annotated. Converbs are annotated according to their function and the finite verbs are separated into two groups: predicates of type 1 contain complex movements (a movement expressed by more than one verb) and phase verbs (e.g. begin), predicates of type 2 contain verbs with aspectual meaning and auxiliaries.

Table 21 *Types of converbs*

tag	**description**
adv	adverbial
s:coord	coordinated
s:sub	subordinated
cvb:pred	converbal predicate

Table 22 *Finite Verb form*

tag	description
v:pred1	complex movements, phase verbs
v:pred2	aspectual meaning, auxiliaries

(18)	**ref**	PMP_1961_BodylessHead_flk.096			
	ts	Ne:jqum pelgalɨk warkl'e übəraŋ.			
	tx	ne:jqum	pelgalɨk	warkl'e	übəraŋ.
	mb	ne:-j-qum	pel-galɨk	wark-l'e	übə-r-a-ŋ
	mp	ne:-l-qum	pelə-galɨk	wargɨ-le	übə-r-ɨ-ŋ
	ge	woman-ADJZ-person.[SG.NOM]	friend-CAR	live-CVB	begin-FRQ-EP-3SG.S
	gr	женщина-ADJZ-человек.[SG.NOM]	друг-CAR	жить-CVB	начать-FRQ-EP-3SG.S
	mc	n-n>adj-n-n:num-case	n-n>adv	v-v>adv	v-v>v-infl:ins-v:pn
	ps	N	N	ADV	V
	SyF	np.h:S			v:pred
	CVB			cvb:pred	v:pred1
	fr	Женщина одна стала жить.			
	fe	The woman begins to live on her own.			
	fg	Die Frau beginnt allein zu leben.			

References

Brykina, Maria, Gusev, Valentin, Szeverényi, Sándor & Wagner-Nagy, Beáta (2016): Nganasan *Spoken Language Corpus (NSLC)*. Archived in Hamburger Zentrum für Sprachkorpora. Version 0.1. Publication date 2016-12-23. http://hdl.handle.net/11022/0000-0001-B36C-C.

Brykina, Maria, Gusev, Valentin, Szeverényi, Sándor & Beáta Wagner-Nagy (2018): *Nganasan Spoken Language Corpus (NSLC)*. Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 12.06.2018. http://hdl.handle.net/11022/0000-0007-C6F2-8

Götze, Michael et al. (2007): Information structure. In: Dipper, S., Götze, M. &S. Skopeteas (eds.): *Information Structure in Cross-Linguistic Corpora.* Interdisciplinary Studies on Information Structure 07 (2007): 147-187, Available online at http://edoc.hu-berlin.de/oa/reports/reQ5PntJcwYs/PDF/23TFAo8H6FW2.pdf [Accessed: 5.2.2013]

Haig, Geoffrey and Stefan Schnell (2014): *Annotations using GRAID (Grammatical relations and animacy in discourse). Introduction and guidelines for annotators*. Version 7.0, Available online at https://www.uni-bamberg.de/fileadmin/aspra/Publications/GRAID7.0_manual.pdf, [Accessed: 03.07.2016]

Helimski, Eugen (1998): Selkup. In: Abondolo, Daniel (ed.). *The Uralic Languages*. London – New York: Routledge, 548–579.

Glushkov, Sergej, Bajdak, Alexandra & Natalya Maksimova (2013): Диалекты селькупского языка. In: Tuchkova, Natalya et al. (eds.). *Селькупы. Очерки традиционной культуры и селькупского языка*. Tomsk, 49–63.

Riester, Arndt & Stefan Baumann (2014): *RefLex Scheme – Annotation Guidelines.* http://www.ims.uni-stuttgart.de/institut/mitarbeiter/arndt/doc/RefLex-guidelines-01aug-2014.pdf.

Russian Census (2010): *Всероссийская перепись населения* 2010. Том 4. Национальный состав и владение. http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm

Wagner-Nagy, Beáta & Sándor Szeverényi & Valentin Gusev (2018): *User's Guide to Nganasan Spoken Language Corpus*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology. Vol. 1. https://doi.org/10.14232/wpcl.2018.1 .

Appendix

Appendix 1: Characters used in the corpus

Corpus	IPA	Cyrillic (original source)	Corpus	IPA	Cyrillic (original source)
a	a	а	h	h	х, ӽ
ä	æ	ä	j	j	й
e	e	е	k	k	к
ɛ	ɛ	э	l	l	л, l
ə	ə	ӭ, ъ	m	m	м
i	i	и, и̇, і	n	n	н
ɨ	ɨ	ы	ŋ	ŋ	ң, нг
ɪ	ɪ	no Cyrillic source	p	p	п
o	o	о	q	q, G	қ, k, г̦
ɔ	ɔ	no Cyrillic source	r	r	р
ö	ø	ö	s	s	с
u	u	у	š	ʃ^j	ш
ü	y	ӱ	t	t	т
č	t͡ʃ	ч	w	v, β	в
d	d	д	z	z	з, ц
g	g	г	ʒ	ʒ	ж
ɣ	ɣ	Y	ǯ	d͡ʒ	җ, дж

Appendix 2: Abbreviations

1	first person	DUR	durative	PROP	proprietive
2	second person	EMPH	emphatic clitic	PST	past
3	third person	EP	epenthetic vowel	PTCP	participle
ABL	ablative	EX	existential verb	REP	reportative
ABST	abstract noun from verb	FRQ	frequentative	RES	resultative
ACC	accusative	FUT	future	RFL	reflexive
ACT	nomen actionis	GEN	gentive	s	subjective conjugation
ADJZ	adjectivizer	HAB	habituative	SG	singular
ADV	adverb	ILL	illative	SING	singulative
ALL	allative	IMP	imperative	SUB	subjunctive mood
AN	animate	INCH	inchoative	TR	transitive
AOR	aorist	INDEF	indefinitive	TRL	translative
ATTEN	attenuative	INF	infinitive	US	usative
CAP	captative	INFER	inferential	VBLZ	verbalizer
CAR	caritive	INSTR	instrumental		
CAUS	causative	INT.PF	intensive perfective		
COM	comitative	IPFV	imperfective suffix		
COND	conditional	ITER	iterative		
CONJ	conjunctive	LOC	locative		
COR	coordinative	MULTS	multisubjective		
CRC	connective reciproc	NEG	negative marker		
CVB	converb	o	objective conjugation		
DAT	dative	OBL	oblique case		

DES	desiderative	OPT	optative		
DETR	detransitive	ORD	ordinal numeral forming suffix		
DIM	diminutive	PL	plural		
DRV	derivational suffix	PREV	preverb		
DU	dual	PROL	prolative		

Appendix 3: Research based on the corpus

Behnke, Anja (2021): *Syntaktische Strukturen im Selkupischen. Eine korpusbasierte Untersuchung der zentralen und südlichen Dialekte*. Logos Verlag Berlin.

Behnke, Anja (2020): Converbal constructions in Selkup. *Eesti ja soome-ugri keeleteaduse ajakiri/Journal of Estonian and Finno-Ugric Linguistics*. 12–2: 137–165.

Budzisch, Josefina (2021): *Definitheit im Selkupischen*. (Studia Uralo-Altaica 55). Szeged.

Budzisch, Josefina (2021): Marking strategies of attributive possession in Selkup: A study of frequency and types of possession. *Finnisch-Ugrische Forschungen*. (to appear)

Budzisch, Josefina (2017): On the non-possessive use of possessive suffixes in Central and Southern Selkup. *Ural-Altaic Studies Scientific Journal* 25: 58–67.

Budzisch, Josefina (2017): Existential, locative and possessive sentences in Selkup dialects. *Finnisch-Ugrische Mitteilungen* 41: 45–62.

Budzisch, Josefina (2015): Possession in Southern Selkup. *Tomsk Journal of Linguistics and Anthropology* 4: 45–50.

Budzisch, Josefina & Ulrike Kahrs (2020): Cardinal directions in Selkup. *Finnisch-Ugrische Mitteilungen* 44: 1–19.

Gusev, Valentin (2017): On the etymology of Auditive in Samoyedic. *Finnisch-Ugrische Mitteilungen* 41: 131–152.

Harder, Anja (2020): *Syntaktische Strukturen im Selkupischen. Eine korpusbasierte Untersuchung der zentralen und südlichen Dialekte*. Dissertation, Universität Hamburg.

Harder, Anja (2017): Grammaticalization of spatial expressions in Central and Southern Selkup. *Finnisch-Ugrische Mitteilungen* 41: 153–174.

Wagner-Nagy, Beáta & Alexandre Arkhipov (2020): Comitative constructions in Nganasan and Selkup. In: Bartens, H.-H.; Larsson, L.-G.; Mattsson, K.; Molnár, J. & Savolainen, T. (eds.). *Kīel joug om šīld*: Festschrift zum 65. Geburtstag von Eberhard Winkler. Wiesbaden: Harrassowitz, 425–441. (with Alexandre Arkhipov).

Wagner-Nagy, Beáta (2020) Predicative possessive constructions in Selkup dialects. In: Dalmi, Gréte, Witkos, Jacek & Piotr Ceglowski (eds.): *Approaches to Predicative Possession*. The View from Slavic and Finno-Ugric. London: Bloomsburry, 205–220.

Wagner-Nagy, Beáta (2017): The essive-translative in Selkup and Kamas. In: De Groot, Caspar (ed.): *Uralic Essive and the Expression of Impermanent State* (Typological Studies in Language119). Amsterdam: John Benjamins, 479–495.

Wagner-Nagy, Beáta (2017): Recipient marking in the Samoyedic languages. In: Krause, Arne, Lehmann, Gesa, Thielmann, Winfried & Caroline Trautmann (eds.): *Form und Funktion.* Festschrift für Angelika Reeder zum 65. Geburtstag. Tübingen: Stauffenburg Verlag, 259–271.

Wegener, Hannah (2018): On differential object marking in Southern and Central Selkup. *Journal of Estonian and Finno-Ugric Linguistics* 9: 169–186.

Appendix 4: Text sources

Bajdak, Alexandra – Nadezhda Fedotova – Natalya Maksimova (2010): Селькупские тексты. In: Filchenko, Andrey (ed.)*: Annotated Folklore Prose Texts of Ob-Yenissey Language Area.* Tomsk: Veter, 133–184.

Bajdak, Alexandra – Natalya Maksimova (2002): *Дидактизация оригинального текста: селькупский язык.* Tomsk.

Bajdak, Alexandra – Natalya Maksimova (2009): Селькупский текст. In: Filchenko, Andrey (ed.): *Annotated Folk and Daily Prose Texts in the Languages of Ob-Yenissei Linguistic Area*. Tomsk: Veter, 45–90.

Bajdak, Alexandra – Natalya Maksimova (2012): Селькупские тексты. In: Filchenko, Andrey (ed.): *Annotated Folk and Daily Prose Texts in the Languages of Ob-Yenissei Linguistic Area*. Tomsk: Veter, 72–100.

Bajdak, Alexandra – Natalya Maksimova (2013): Селькупские тексты. In: Filchenko, Andrey (ed.): *Annotated Folk and Daily Prose Texts in the Languages of Ob-Yenissei Linguistic Area*. Tomsk: Veter, 153–201.

Bajdak, Alexandra – Natalya Maksimova (2015): Селькупские тексты. In: Filchenko, Andrey (ed.): *Annotated Folk and Daily Prose Texts in the Languages of Ob-Yenissei Linguistic Area*. Tomsk: Veter, 108–149.

Bajdak, Alexandra – Natalya Tuchkova (2004): Эписоды "Эпоса об Итте" в чулылькупском диалектном ареале”. In: *Коренные народы Сибири: проблемы историографии, истории, этнографии, лингвистики.* Tomsk, 51–64.

Bekker, Erika (1978): *Категория падежа в селькупском языке*. Tomsk: Издательство Томского университета.

Bekker, Erika (1980): Селькупские тексты. In: *Сказки народов Сибирского Севера*. Tomsk: Издательство Томского университета, 55–71.

Bykonia, Valentina – Alexandra Kim – Shimon Kuper – Natalya Maksimova – I. Ilyashenko (1999): *Сказки нарымских селькупов.* Tomsk: NTL.

Castrén, Matthias (1855): *Nordische Reisen und Forschungen* 8. St. Petersburg: Kaiserliche Akademie der Wissenschaften.

Dulzon, Andreas (1966a): *Кетские сказки.* Tomsk: Издательство Томского университета.

Dulzon, Andreas (1966b): Селькупские сказки. In: *Языки и топонимия Сибири. Том 1*. Tomsk, 96–158.

Grigorovskij, Nikolaj (1879): *Азбука сюссогой гулани.* Kazan: Типография Императорского Университета.

Grigorovskij, Nikolaj (1883): *Итя (Сказка обских самоедов)*. Томские губернские ведомости 24.

Janhunen, Juha (1975): *Etä-sukukielet. Lapp, Volga-Finnic, Permian-Finnic, Ob-Ugrian, Samoyed.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Katz, Hartmut (1975): *Materialien vom Tym*. Selcupica 1. München.

Katz, Hartmut (1988): *Die Märchen in Grigorovskis Azbuka. Transkription, Übersetzung, Kommentar*. Selcupica 4. München.

Kim, Alexandra (2002): К истории изучения южноселькупского фольклора: тымские материалы Л. Сабо. In: *Образование в сибири: актуальные проблемы истории и современность*. Tomsk, 204–206.

Korobeynikova, Irina (2014): *Сказки и рассказы селькупки ирины*. Tomsk: Veter.

Künnap, Ago (1992): *Selkupin Iidjä-satu Tšajan murteela vuodelta 1913*. Fenno-ugristica 18, 141–147.

Kuzmina, Angelina (1967): Диалектологические материалы по секупскому языку. In: *Исследования по языку ифольклору. Вып. 2.* Novosibirsk, 267–329.

Kuznecova, A. – Olga Kazakevich – L. Ioffe – Eugen Helimski (1993): *Очерки по селькупски языку.Тазовский диалект.* Том 2. Moscow.

Lehtisalo, Toivo (1960). *Samojedische Sprachmaterialien*. Mémoires de la Société Finno-Ougrienne 122. Helsinki.

Morev, Yuri – Nina Denning – Valentina Bykonia – G. Mikhenina – Alexandra Kim (1981): Самодийские тексты. Селькупские тексты. In: *Сказки народов Сибирского севера*. Tomsk: Издательство Томского университета, 122–143.

Сказки чаинских селькупов [Skazki chainskih selkupov] (2001): In: *Земля чаинская: сборник научно-популярных очерков 100-летию села Подгорного*. Tomsk, 115–126.

Szabó, Lászlo (1966): Szelkup szövegek szójegyzékkel (Tymi nyelvjárás). *Nyelvtudományi Közlemények* 68, 266–277.

Szabó, Lászlo (1967): *Selkup texts with phonetic introduction and vocabulary*. Uralic and Altaic Series 75. Indiana: Bloomington.

Tuchkova, Natalya (2002): К вопросу о педагогических традициях селькупов. In: *Образование в сибири: актуальные проблемы истории и современность*. Tomsk, 195–204.

Tuchkova, Natalya – Eugen Helimski (2010): *О материалах А. И. Кузьминой по селькупскому языку*. Hamburg: Institut für Finnougristik/Uralistik.

Tuchkova, Natalya – Beáta Wagner-Nagy (2015): *„sēľďe nūn qǭdi īťťe...“. «Семи богов мудростью обладающий Итте...» Тексты с героем Итя в селькупском фольклоре Часть 1. Итя-тексты.* Tomsk: Томский государственный педагогический университет.